Received: 15 November 2021 | Revised: 18 January 2022 | Accepted: 20 January 2022

DOI: 10.1002/nbm.4702

RESEARCH ARTICLE

NMR IN BIOMEDICINE WILEY

Comparison of seven modelling algorithms for γ-aminobutyric acid–edited proton magnetic resonance spectroscopy

Alexander R. Craven[1,2,3] | Pallab K. Bhattacharyya[4] | William T. Clarke[5,6] | Ulrike Dydak[7] | Richard A. E. Edden[8,9] | Lars Ersland[1,2] | Pravat K. Mandal[10,11] | Mark Mikkelsen[8,9,12] | James B. Murdoch | Jamie Near[13,14,15] | Reuben Rideaux[16] | Deepika Shukla[10,17,18] | Min Wang[19] | Martin Wilson[20] | Helge J. Zöllner[8,9] | Kenneth Hugdahl[1,21,22] | Georg Oeltzschner[8,9]

[1]Department of Biological and Medical Psychology, University of Bergen, Bergen, Norway

[2]Department of Clinical Engineering, Haukeland University Hospital, Bergen, Norway

[3]NORMENT Center of Excellence, Haukeland University Hospital, Bergen, Norway

[4]Cleveland Clinic Foundation, Imaging Institute, Cleveland, Ohio, USA

[5]Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK

[6]MRC Brain Network Dynamics Unit, University of Oxford, Oxford, UK

[7]School of Health Sciences, Purdue University, Indiana, West Lafayette, USA

[8]Russell H. Morgan Department of Radiology and Radiological Science, The Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

[9]F. M. Kirby Research Center for Functional Brain Imaging, Kennedy Krieger Institute, Baltimore, Maryland, USA

[10]NeuroImaging and NeuroSpectroscopy (NINS) Laboratory, National Brain Research Centre, Gurgaon, India

[11]Florey Institute of Neuroscience and Mental Health, Parkville, Melbourne, Victoria, Australia

[12]Department of Radiology, Weill Cornell Medicine, New York, New York, USA

[13]Centre d'Imagerie Cérébrale, Douglas Mental Health University Institute, Montreal, Canada

[14]Department of Biomedical Engineering, McGill University, Montreal, Canada

[15]Department of Psychiatry, McGill University, Montreal, Canada

[16]Queensland Brain Institute, The University of Queensland, Brisbane, Australia

[17]Perinatal Trials Unit Foundation, Bengaluru, India

[18]Centre for Perinatal Neuroscience, Imperial College London, London, UK

[19]College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China

[20]Centre for Human Brain Health and School of Psychology, University of Birmingham, Birmingham, UK

[21]Division of Psychiatry, Haukeland University Hospital, Bergen, Norway

[22]Department of Radiology, Haukeland University Hospital, Bergen, Norway

Abbreviations used: Cho, choline; CI, confidence interval; Cr, creatine; CRLB, Cramér–Rao lower bound for uncertainty; diff, difference (edited) spectrum; ECC, eddy-current correction; FD, frequency domain; FID, free induction decay (observed time-domain signal); FWHM (linewidth), full width at half maximum; GABA, γ-aminobutyric acid; GABA+, total edited signal at 3 ppm (GABA with underlying coedited signal); Gln, glutamine; Glu, glutamate; Glx, combined signal of glutamate + glutamine; GSH, glutathione; H_2O ($_{noTC}$), water (referenced without tissue class correction); HSVD, Hankel singular value decomposition; i.u., institutional units; ICC, intraclass correlation coefficient; LCM, linear combination modelling; MAD, median absolute deviation; MEGA-PRESS, Mescher–Garwood point-resolved spectroscopy; MMx(y), macromolecule signal around x(.y) ppm; NAA, N-acetylaspartate; NAAG, N-acetylaspartylglutamate; p_{holm}, Holm–Bonferroni adjusted *p* value; ppm, parts per million; Q-Q, quantile-quantile; R1–4, adopted rejection criteria; SD, standard deviation; SNR, signal-to-noise ratio; tCr, total creatine (creatine +phosphocreatine); TD, time domain; VPC, variance partition coefficient.

Correspondence
Alexander R. Craven, Department of Biological and Medical Psychology, University of Bergen, Jonas Lies vei 91, 5009, Bergen, Norway.
Email: alex.craven@uib.no

Funding information
H2020 European Research Council Advanced, Grant/Award Number: #693124; National Institutes of Health (NIH), Grant/Award Numbers: K99/R00 AG062230, S10 OD021648, P41 EB031771, P41 EB015909, R01 EB016089, R01 EB023963, R01 EB028259, R21 HD100869; Wellcome Trust and the Royal Society, Grant/Award Number: 102584/Z/13/Z; India–Australia Strategic Biotechnology Funding, Grant/Award Number: BT/Indo-Aus/10/31/2016/PKM; Australian Research Council, Grant/Award Number: DE210100790

Edited MRS sequences are widely used for studying γ-aminobutyric acid (GABA) in the human brain. Several algorithms are available for modelling these data, deriving metabolite concentration estimates through peak fitting or a linear combination of basis spectra. The present study compares seven such algorithms, using data obtained in a large multisite study. GABA-edited (GABA+, TE = 68 ms MEGA-PRESS) data from 222 subjects at 20 sites were processed via a standardised pipeline, before modelling with FSL-MRS, Gannet, AMARES, QUEST, LCModel, Osprey and Tarquin, using standardised vendor-specific basis sets (for GE, Philips and Siemens) where appropriate. After referencing metabolite estimates (to water or creatine), systematic differences in scale were observed between datasets acquired on different vendors' hardware, presenting across algorithms. Scale differences across algorithms were also observed. Using the correlation between metabolite estimates and voxel tissue fraction as a benchmark, most algorithms were found to be similarly effective in detecting differences in GABA+. An interclass correlation across all algorithms showed single-rater consistency for GABA+ estimates of around 0.38, indicating moderate agreement. Upon inclusion of a basis set component explicitly modelling the macromolecule signal underlying the observed 3.0 ppm GABA peaks, single-rater consistency improved to 0.44. Correlation between discrete pairs of algorithms varied, and was concerningly weak in some cases. Our findings highlight the need for consensus on appropriate modelling parameters across different algorithms, and for detailed reporting of the parameters adopted in individual studies to ensure reproducibility and meaningful comparison of outcomes between different studies.

KEYWORDS
GABA, macromolecule, MEGA-PRESS, MRS, quantification, spectral editing

1 | INTRODUCTION

Several software packages and modelling algorithms are available for processing and quantifying MR spectroscopy (MRS) data. While they are all designed to extract quantitative estimates of metabolite levels from spectra, the packages differ significantly in their approach to processing and modelling the underlying data, and isolating the components of interest from any artefactual signals therein. This may give rise to systematic differences in metabolite estimates between different software packages. While an effect of 'choice of software' has been documented for short-echo-time data,[1–4] similar studies for γ-aminobutyric acid (GABA)–edited MRS quantification are lacking.

Spectral editing experiments,[5,6] such as the widely used Mescher–Garwood point-resolved spectroscopy (MEGA-PRESS) technique for the selective detection of GABA, present a special case for quantification. In a typical MEGA-PRESS editing sequence, two interleaved subspectra are acquired: the edit-ON subspectrum, in which coupling to GABA spins at three parts per million (ppm) is refocused, and the edit-OFF subspectrum, in which it is not. Subtracting the edit-ON and edit-OFF subspectra yields a relatively sparse difference spectrum, featuring prominent broad signal for GABA (with underlying macromolecule contributions) at 3 ppm and coedited signals including glutamate (Glu) and glutamine (Gln) peaks (usually reported collectively as the combined signal of glutamate + glutamine [Glx]) around 3.75 ppm, and strong negative peaks close to the editing frequency — primarily N-acetylaspartate (NAA) and N-acetylaspartylglutamate (NAAG).

Most notable among the challenges for modelling edited spectra are coedited macromolecular signals coupled to spins near the editing frequency,[6] some of which appear in the same frequency range as the GABA and Glx signals and therefore interfere with their unambiguous modelling. As they are broad and poorly characterised, no consensus currently exists on how they should be accounted for in the modelling stage. Constrained by the inability to reliably separate GABA and macromolecules, their composite is commonly reported: total edited signal at 3 ppm (GABA with underlying coedited signals [GABA+]).

A rigorous assessment of the comparability of GABA+ estimates obtained across a range of different analysis software packages is currently lacking. Several prior studies[7–10] have investigated the test–retest reproducibility of GABA+ estimates using a small selection of available software packages, but without detailed examination of the differences in estimates arising between software packages. Each considered data from a single site only. Another study[11] has investigated GABA+ estimates from Gannet and Tarquin compared with a simulated 'ground truth', specifically with respect to the influence of signal-to-noise ratio (SNR) and linewidth on estimates, showing that the two algorithms agreed under favourable conditions of linewidth and SNR but diverged under poorer conditions; however, only two algorithms were included in this analysis. A recent conference paper[12] has reported early findings from GABA-edited MEGA-PRESS data showing moderate associations between five different algorithms, with data from four sites (representing two scanner vendors), albeit with divergent processing. A more thorough examination, covering a broader range of sites and an extended selection of contemporary algorithms, is required to better characterise the noted discrepancies.

Therefore, to establish the degree to which different software packages agree in estimating GABA+ from MEGA-PRESS data, this study compares GABA+ estimates from seven modelling algorithms: FSL-MRS,[13] Gannet,[14] LCModel,[15] Osprey,[16] Tarquin,[17,18] AMARES[19] and QUEST,[20,21] with the last two implemented in the jMRUI software package.[22,23] Estimates of Glx from the difference spectra are also considered. Detailed characterisation of differences observable across algorithms is essential for meaningful comparison of findings reported from different tools, and particularly in reconciling any discrepancies therein.

2 | METHODS

2.1 | Data

Data from 20 3-T MRI scanners from the three major manufacturers (GE, Philips and Siemens), each at a different site, were obtained from the Big GABA[24,25] repository on NITRC (https://www.nitrc.org/projects/biggaba). GABA-edited spectra (TR/TE = 2000/68 ms, 320 averages, editing at 1.9/7.46 ppm for edit-ON/-OFF, respectively) and corresponding water-unsuppressed reference data (eight or 16 averages) were obtained from a 3 x 3 x 3 cm^3 voxel in the posterior cingulate region, from 222 consenting adult volunteers (aged 18–36 years, approximately an even female/male split, having no known neurological or psychiatric illness), in accordance with ethical standards of their respective local institutional review boards. Subjects consented to the sharing of anonymised data, with allowance for further study. Datasets also included T_1-weighted structural MR images, which were used for tissue segmentation.

This extensive collection of datasets was acquired in an international collaborative study; several aspects have been previously reported,[24–26] with a focus on comparability across sites and vendors. Full details on the acquisition protocol, software and hardware configurations and sample composition may be found in these papers, and are summarised in Table S1. Detailed vendor-specific parameters have been reported previously.[25,27]

2.2 | Processing

To the maximum extent practical, data were prepared for each algorithm using a common pipeline, to avoid variations in processing that might otherwise confound observations regarding the model fit. Original data in vendor-specific format were imported using the GannetLoad function from Gannet (v. 3.1). The GannetLoad function was chosen because it had the broadest support for the diverse file formats and sequence implementations present in these datasets. This function applies coil combination where necessary, and initial categorisation of individual free induction decay (FID) observed time-domain signals into edit-ON/OFF subspectra and water reference spectra. Although GannetLoad also incorporates a full processing pipeline, we did not make use of this, instead electing to implement a generalised pipeline in accordance with current consensus recommendations,[28] using the processing tools from FID-A.[29] The rationale for this was to provide a common, neutral starting point for quantification across all the algorithms to be assessed, rather than one which may have been tuned for a particular quantification algorithm. Additionally, the standard Gannet pipeline performs line-broadening and zero-filling, which invalidates assumptions for error calculations in linear-combination modelling algorithms such as LCModel.

Initially, motion-corrupted transients were removed by comparing the root mean square of the difference between each transient and the median of transients in the time domain, rejecting those which differed from the mean by more than four standard deviations (SDs). This unlikeness metric was calculated independently within the edit-ON and edit-OFF subspectra but applied pairwise. To correct for frequency and phase drift, the remaining FIDs were then aligned in the frequency domain using the spectral registration method,[30] iteratively on a variable, restricted frequency range (~1.6–5.5 ppm in the first iteration, reducing to ~1.8–4.0 ppm in subsequent iterations), before averaging within the edit-ON and edit-OFF subspectra. Eddy-current correction (ECC) was applied,[31] before zero-order phase adjustment of each subspectrum according to a dual-Lorentzian model for creatine (Cr) and choline (Cho) defined in Gannet[14] and implemented in Osprey. Thereafter, edit-ON and edit-OFF

subspectra were aligned by spectral registration,[30] and sum and difference spectra were calculated arithmetically on the time domain data, dividing by the number of subspectra. All resultant spectra and subspectra were frequency-shifted such that the main Cr peak (from the dual-Lorentzian model for Cr and Cho) in the sum spectrum appeared at 3.027 ppm. After calculation of the difference spectrum, residual water was filtered from edit-ON and edit-OFF data separately with the Hankel singular value decomposition (HSVD) method.[32] A final automated check was performed to ensure correct ON/OFF ordering and orientation of all resultant spectra, flipping where necessary. Processed data were exported with the same resolution (number of samples and sweep width/dwell time) as the incoming data, without line-broadening or zero-filling. Processing and modelling workflow are summarised in Figure 1.

2.2.1 | Quality control: Processing

Processed spectra were tested against two rejection criteria, designated **R1** and **R2** in subsequent usage:

- R1 (R1–4, adopted rejection criteria) captures spectra having strongly aberrant features in the fit range: processing was deemed to have failed if the 0-lag cross-correlation of the normalised, reconstructed frequency domain difference spectrum in the metabolite range (2.6–4.2 ppm) with the normalised mean of all other difference spectra was less than 0.5 or differed from the group mean by more than three SDs.
- R2 establishes thresholds on basic signal quality metrics: SNR (< 80, defined by maximum peak height around NAA_{diff} in the [1.8, 2.2] ppm interval, over SD across the [−2, 0] ppm range) and linewidth (full width at half maximum [FWHM] > 10 Hz,[33]) measured from NAA_{diff}.

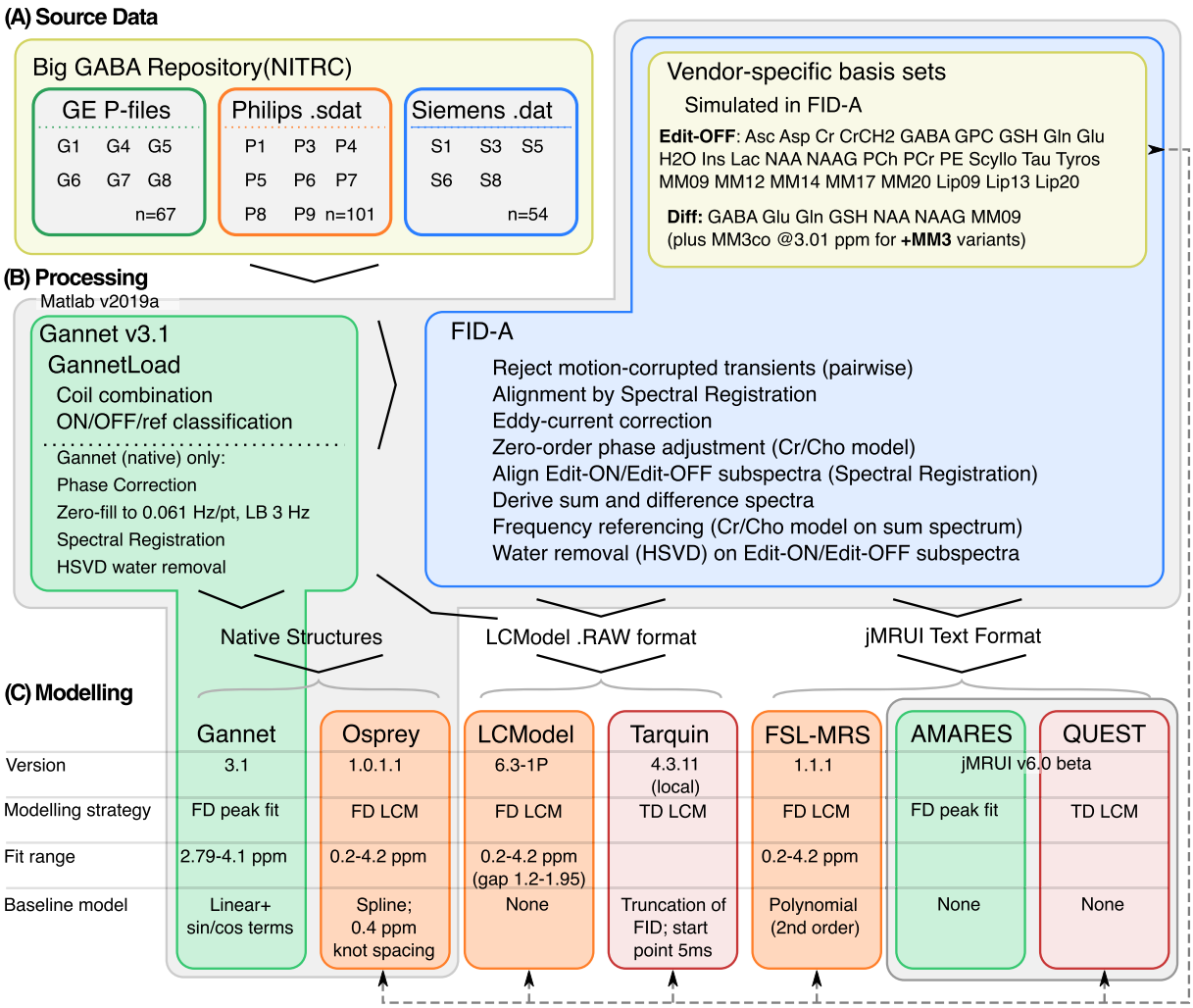

FIGURE 1 (A) Source data, (B) Processing and (C) Modelling workflow, summarising key differences between the algorithms assessed. Cho, choline; Cr, creatine; Diff, difference (edited) spectrum; FD, frequency domain; LCM, linear combination modelling; TD, time domain

Data deemed to have failed at the processing stage were still passed to the fit algorithms but flagged as having failed and excluded from evaluation of groupwise statistics (such as median estimates) in further analysis.

2.3 | Initial fit and quantification

Identically processed data were fed into each algorithm. To the maximum extent practical, data were modelled using the developer-supplied default or recommended configuration parameters for GABA-edited MEGA-PRESS data, to yield outcomes representative of those which researchers could expect without extensive local optimisation.

Batch processing for all algorithms was automated in Matlab (v. 2019a), with the exception of the jMRUI-based algorithms, for which processed data were exported then processed as batches (grouped by manufacturer and spectral resolution) in a standardised but manual procedure through the jMRUI user interface. As the commonly used default processing pipeline for Gannet incorporates zero-fill and line-broadening factors not present in the standardised pipeline adopted here, we report outcomes both from the standardised processing pipeline (hereafter denoted 'Gannet'), and from data processed with Gannet's own default pipeline, denoted 'Gannet (native)'. Tarquin fitting is often performed with an internally simulated basis set; we also assess outcomes from this mode of operation, hereafter denoted 'Tarquin (internal)'.

Full details on the operation of each method are supplied in the supporting information, section C, for quantification of the edited difference spectra. To facilitate concentration scaling to an internal Cr reference, corresponding edit-OFF subspectra are also modelled; this is described in the supporting information, section E. Concentration estimates are reported both relative to total creatine (tCr), and with respect to an internal water reference; the complexities and relative merits of each approach are described in Near et al.[28]

As the specifics of each algorithm's water referencing procedure varied considerably, scaling as documented for the respective algorithms was first reversed to yield a raw ratio of signal intensities, before applying tissue-class correction[34] using previously derived tissue fractions.[24] Full details on the adjustment for each algorithm are provided in the supporting information, section D. Water-scaled, tissue-class–corrected molar concentration estimates are hereafter denoted '/H_2O'; concentration estimates scaled to water with no adjustment for tissue class (assuming pure water concentration as per Equation (3) of [35]) are also calculated, denoted '/H_2O_{noTC}'.

2.3.1 | Basis set preparation and prior knowledge

All the algorithms examined require some degree of prior knowledge to describe expected spectral features, either in the form of parameter constraints or simulated basis sets. In the present study, prior knowledge was standardised as far as possible: all algorithms requiring a basis set were supplied with the same simulated basis set appropriate to the dataset, while both algorithms parameterising individual peaks (Gannet and AMARES) were supplied with similar model parameters.

For comparison of the basis set algorithms (FSL-MRS, LCModel, Osprey, QUEST and Tarquin), a standard simulated basis set specific to each hardware vendor was adopted. As a starting point, vendor-specific basis sets for GABA-edited MEGA-PRESS (TE = 68 ms) that are distributed with Osprey were used; these are derived from fast spatially resolved 2D density-matrix simulations[36] implemented in FID-A using ideal excitation pulses and vendor-specific refocusing pulses and timings, and using chemical shifts and J-coupling coefficients from Kaiser et al.[37] These incorporated metabolite basis functions for GABA, Glu, Gln, glutathione (GSH), NAA, NAAG and a Gaussian component (FWHM = 10.9 Hz) representing coedited macromolecules around 0.91 ppm (macromolecule signal around x(.y) ppm [MM09ex]).

A variation of this basis set was created, incorporating an additional Gaussian component at 3.0 ppm (simulated with FWHM = 14 Hz and scaled intensity equivalent to two protons) to represent coedited macromolecule signal underlying the GABA peak around 3.0 ppm. The amplitude scaling is consistent with the assumptions of a pseudo-doublet GABA signal at 3 ppm, and around 50% macromolecule contribution[38–42] to the observed signal in that area; the FWHM parameter had been optimised previously[43] on an aggregate subset of the data, over the 1–20 Hz range. This component, denoted MM3co, allowed the influence of macromolecule modelling on the various algorithms to be examined; subsequent use of this basis set is annotated with +MM3. The interaction of this component with baseline stiffness and soft constraint models similar to those of[43–45] is explored for Osprey and LCModel in the supporting information, section B. All basis set algorithms were run both with and without the MM3 component; in all cases, the reported GABA+ values include contributions from the underlying macromolecule signal, either explicitly in cases where the MM3 component was modelled (i.e., GABA + MM3co), or implicitly in cases where it was not.

2.3.2 | Quality control: Modelling

The available quality metrics vary between algorithms; all except Osprey report some form of modelling uncertainty (% SD, % Cramér-Rao lower bound for uncertainty [CRLB] of metabolite estimates or % fit error for the model), and most report SNR and linewidth of water and/or some

metabolite components. As the specifics of each algorithm's SNR and linewidth calculation vary, independently derived values are assessed at the processing stage (R2; section 2.2.1). Adopting rather liberal criteria, individual fits were flagged as having failed if either of the following additional criteria were met; in cases where a given metric was not available, the condition was ignored. The criteria below are designated **R3** and **R4** for subsequent usage.

- R3: %SD, CRLB or FitError for GABA+, Glx_{diff} or tCr_{edit_off} estimate exceeded 50% (as per,[33,46] acknowledging that this strategy must be used with caution[47]).
- R4: Final, scaled estimate for any target metabolite (GABA+/H_2O, Glx_{diff}/H_2O, GABA+/tCr_{edit_off}, $Glx_{diff}/tCr_{edit\_off}$) differing from the median value by more than five times the median absolute deviation (MAD)[48] for that algorithm; this was intended to capture any poor fits not flagged by any other criteria.

Visual inspection of data, fit outcomes and residuals was also performed, to confirm that no grossly aberrant outcomes eluded the defined rejection criteria. All subsequent analyses are performed after exclusion of individual algorithms' fits (not entire subject datasets) as per these criteria.

2.4 | Statistical analysis of modelling outcomes

After batch modelling, statistical analysis was performed using locally implemented scripts written in Python (v. 3.7.3), using the pandas[49] (v. 0.23.3) data analysis framework, with numeric methods from NumPy,[50] and statistical methods from the SciPy[51] (v. 1.1.0), pingouin[52] and statsmodels[53] (v. 0.12.1) libraries.

Scaled estimates for target metabolite, grouped by algorithm, were tested for normality using the Shapiro–Wilk method,[54] and for comparable distribution of variance between algorithms by the Fligner–Killeen's test,[55] both implemented in SciPy. Limits of agreement between pairs of algorithms were derived, along with their 95% confidence intervals, in accordance with the Bland–Altman method.[56] Estimates grouped by algorithm and manufacturer were compared using Welch's t-test,[57] with Holm–Bonferroni correction[58,59] for multiple comparisons. An adjusted p value of less than 0.05 was considered significant.

An unconditional linear mixed-effects model was fit to water-referenced GABA+ estimates, using R version 3.5.3[60] with the *lme4* package[61] and an implementation derived from.[25] Vendor, site, algorithm and subject factors were incorporated into the fit, to calculate variance partition coefficients (VPCs) estimating the proportion of total variance attributable to each factor. Significance testing was performed using chi-square likelihood ratio tests, against a null hypothesis simulated by parametric bootstrapping (2000 simulations).[62] The Big GABA dataset described herein has previously been assessed with respect to demographics and signal quality.[24]

For each metabolite of interest, a global median was calculated across all subjects and all algorithms. Subsequently, estimates grouped by site and algorithm were linearly scaled to match the global median, thereby removing broad scaling differences observed between certain sites, vendors and algorithms that would otherwise bias inter-algorithm correlations.

The degree of correlation between MRS-detected GABA estimates and voxel tissue fraction has been shown to be an effective index of GABA estimation accuracy,[35] given the differing GABA concentrations between grey and white matter.[63–65] Building on this approach, robust Spearman correlation coefficients between voxel grey matter fraction and GABA+/water referenced without tissue class correction (H_2O_{noTC}) estimates were calculated using the 'skipped' method[66,67] (implemented in pingouin) to exclude bivariate outliers. To determine whether correlation coefficients obtained for any one of the algorithms differed significantly from the correlation obtained across all algorithms, z-scored coefficients were compared (two-tailed), with a significance level of Holm–Bonferroni adjusted p value (p_{holm}) of less than 0.05; similarly, z-scored coefficients were compared between variants without and with the MM3 macromolecule component in the basis set.

Finally, intraclass correlation coefficients (ICCs) were calculated between all algorithms, separately without and with the MM3 component for basis set algorithms, and between pairs of algorithms, using a two-way mixed-effects model for single-rater consistency (ICC (3,1) implemented in pingouin).

3 | RESULTS

3.1 | Fit and residuals

Mean fit outcomes for each algorithm are presented in Figure 2. Note that all basis set algorithms without MM3, except QUEST, show strong residuals in the 3 ppm range; in the case of Osprey (Figure 2F), this appears to have a strong influence on baseline in the vicinity. Variants with MM3 show generally reduced residuals in that range, indicating that the inclusion of a dedicated MM3 basis function leads to more appropriate modelling of the data.

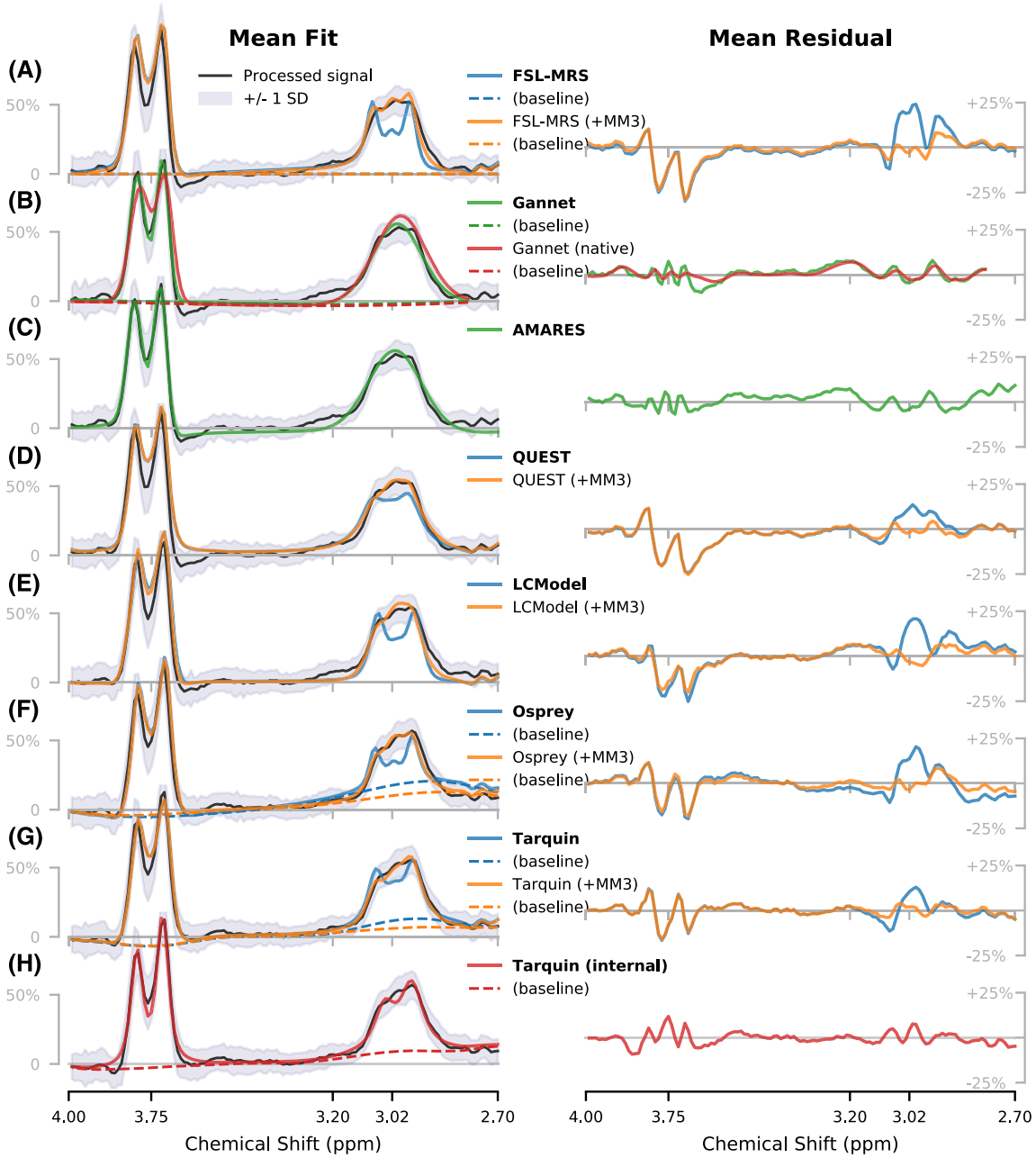

FIGURE 2 Average metabolite and baseline (where applicable) models with corresponding residuals for the GABA+ edited spectra, for each algorithm: (A) FSL-MRS, (B) Gannet, (C) AMARES, (D) QUEST, (E) LCModel, (F) Osprey, (G) Tarquin and (H) Tarquin using its internally-generated basis set. Vertical scaling is normalised; outcomes over the full fit range are presented in Figure S8; outcomes split by vendor are presented in Figure S9

A characteristic hump around 3.2 ppm is handled differently by the various algorithms: peak fitting algorithms Gannet and AMARES (Figure 2B,C) are largely unperturbed, QUEST (Figure 2D) envelopes the entire signal with broader 3.0 ppm peak, while other algorithms fall somewhere in between.

A notable difference between algorithms arises from the differences in baseline estimation practices. While AMARES, QUEST and LCModel do not include a baseline term in their default settings for MEGA-PRESS, and Gannet and FSL-MRS adopt relatively stiff, low-order models, both Tarquin and Osprey attribute a considerable fraction of the edited 3-ppm signal to the baseline. This tendency is mitigated upon the inclusion of the MM3 model.

Finally, there is a distinct pattern to the residuals around the Glx peaks from all basis set algorithms, not present in the fits applying simple peak fitting on a restricted frequency range (Gannet, AMARES).

A summary of basic quality metrics from the fit spectra is presented in Figure S6, along with the number of spectra rejected according to the defined criteria (as per sections 2.2.1 and 2.3.2).

3.2 | Statistical analysis

Shapiro–Wilk testing and subsequent inspection of quantile-quantile (Q-Q) plots indicated that while concentration estimates from most algorithms satisfied the assumption of a normal distribution, several (predominantly Glx_{diff}/tCr estimates) deviated slightly from this. Fligner–Killeen tests revealed mismatched variances between several sets of estimates (predominantly relating to QUEST and LCModel +MM3), which motivated the subsequent adoption of Welch's t-test for groupwise comparisons.

3.2.1 | Water-referenced concentration estimates

Comparisons between algorithms for GABA+/H_2O are summarised in Figure 3, with full details in Table S6, and Bland–Altman plots describing limits of agreement in Figure S10. The global median estimate for GABA+/H_2O across all algorithms and subjects was found to be 3.2 ± 0.4 institutional units (i.u.).

For several quantification algorithms, water-referenced estimates for GABA+/H_2O were found to be significantly higher from Siemens datasets than from other manufacturers, by a factor of 9%–17% (p_{holm} < 0.001, depending on the algorithm) for FSL-MRS (+MM3), Gannet, Gannet (native), AMARES and LCModel, and increased to 41% (p_{holm} < 0.001) for QUEST. Osprey gave significantly lower estimates for Siemens datasets (−28.0%, p_{holm} < 0.001). QUEST (and the +MM3 variant) gave significantly lower estimates for Philips datasets (−16.1% and −7.7%, respectively, p_{holm} < 0.001), and AMARES and Gannet (native) gave lower estimates for GE datasets (−9.5%, p_{holm} < 0.01 and −8.5%, p_{holm} < 0.05, respectively). Median GABA+ estimate across all algorithms was 5.8% higher for Siemens sites (p_{holm} < 0.01). All differences are expressed relative to the mean across all subjects for the respective algorithm. No other variants showed significant effects.

Water-referenced Glx_{diff} estimates from all algorithms were significantly higher for Siemens sites: median Glx_{diff}/H_2O across algorithms +15.7% (p_{holm} < 0.001) relative to group mean. Estimates from Philips sites were somewhat lower (−10.1%, p_{holm} < 0.01).

For data fit without the explicit MM3 component, the unconditional linear mixed-effects model yielded VPCs of 33.8%, 16.4%, 6.4% and 4.0% for algorithm, site, subject and vendor factors, respectively. In this context, the 'subject' factor reflects systematic within-subject variation in estimates, while the residual 39.4% accounts for inherent, systematic between-subject variation, as well as any other variance that could not be accounted for in the model. Parametric bootstrap testing showed all factors to be significant (p_{holm} < 0.001).

3.2.2 | Metabolite-referenced concentration estimates

Estimates for GABA+/tCr_{edit_off} were consistently higher for GE datasets (+17.3% across algorithms, p_{holm} < 0.001) and lower for Siemens datasets (−14.3%, p_{holm} < 0.001). $Glx_{diff}/tCr_{edit\_off}$ ratios were higher in GE datasets (21.9%, p_{holm} < 0.001) and slightly lower in Philips (−5.0%, p_{holm} < 0.05) and Siemens (−6.7%, p_{holm} < 0.05) datasets. As in section 3.2.1, differences are quoted relative to the mean estimate across all

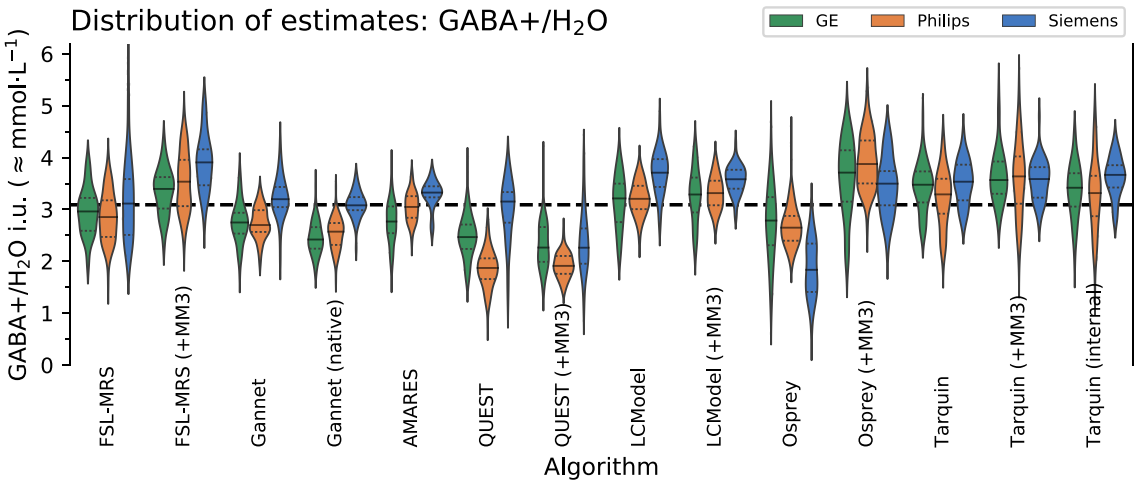

FIGURE 3 Distribution of GABA+/H_2O estimates from each algorithm, grouped by manufacturer. The global median is shown in dashed black

subjects for the respective algorithm. All these trends presented similarly across all modelling algorithms, albeit with varying magnitudes and significance levels.

3.2.3 | Grey matter volume fraction correlation

The relationship between estimated GABA+ and grey matter volume fraction is reported in Figure 4, as an index of estimation accuracy. The accuracy of QUEST (without MM3) was found to be significantly below that of other algorithms ($p_{holm} < 0.01$), while the QUEST +MM3 variant performed comparably with other algorithms. Otherwise, slight differences observable between algorithms were not statistically significant, and no particular trend is evident between algorithm variants without and with MM3 components.

3.2.4 | Correlational analysis

ICC (single-rater consistency) for GABA+ across all algorithms was 0.38 (95% CI 0.32–0.44) without the MM3 component for basis set algorithms, and increased to 0.44 (95% CI 0.39–0.5) with MM3 included, supporting that the inclusion of this dedicated component is

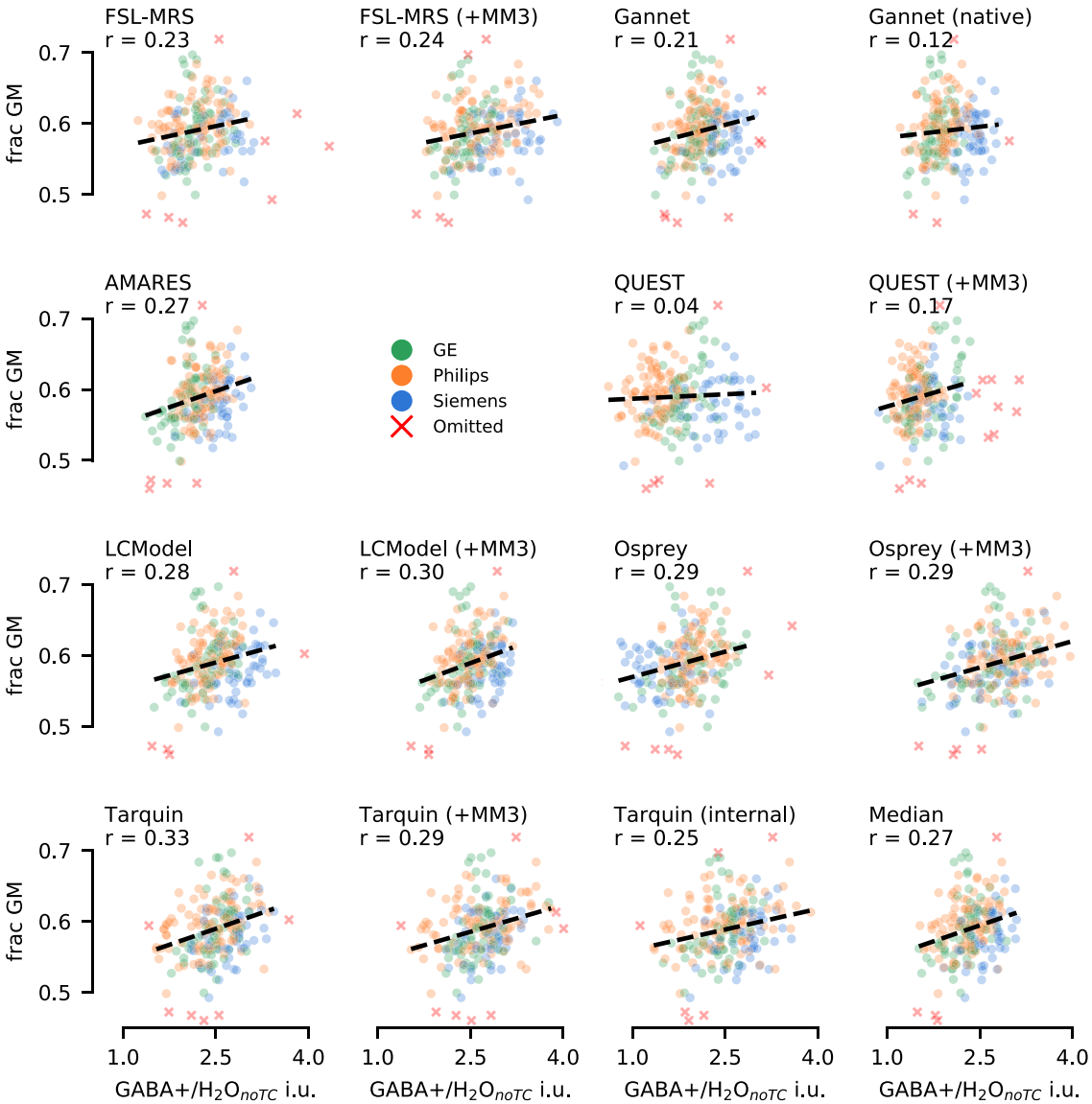

FIGURE 4 Relationship between GABA+ and grey matter (GM), with different modelling strategies for GABA+. Robust (skipped) correlation coefficients are reported, with line-of-best-fit in dashed black

warranted. ICCs between all pairs of algorithms are presented in Figure 5. For fits performed without the MM3 component, GABA+/H_2O estimates showed moderate correlation between most algorithms (typically of the range r = 0.4–0.6; slightly lower when referenced to tCr_{edit_off}). Correlations for AMARES, LCModel and Tarquin were significantly stronger (p_{holm} < 0.01) than the group mean, those for QUEST and Osprey somewhat lower. Inclusion of an MM3 basis set component generally improved concordance with other algorithms for FSL-MRS (p_{holm} < 0.001) and Osprey, the latter at trend level. However, both time domain basis set algorithms (QUEST and Tarquin) showed reduced concordance (at trend level) upon inclusion of the MM3 component.

ICCs for additional metabolites and ratios are presented in Figure S12; Glx_{diff}/H_2O estimates from the edited spectrum correlated more strongly between algorithms (typically of the range r = 0.6–0.8, slightly lower when referenced to tCr_{edit_off}).

4 | DISCUSSION

4.1 | Quality control

Basic signal quality metrics (such as SNR and linewidth) and reliability-of-fit estimates (% CRLB, % fit error) are often used as the basis for rejecting poor fits. However, as seen in Figure S6, these are often not sufficient. While four datasets were deemed to have failed at the processing stage (R1), yielding output barely recognisable as GABA-edited difference spectra, all algorithms 'successfully' fit some of these (Figure S7), with quality metrics that satisfied all other criteria. We therefore repeat the observation that simply filtering results based on these basic signal metrics is inadequate as a means of quality control; the metrics themselves may have limited reliability, particularly in cases where the model does not accurately reflect the experimental data.[68] Consideration must also be given to the shape of the data, fit and residuals themselves, either by algorithmic assessment, or, if feasible, visual inspection.

4.2 | Scale differences between manufacturers

Previous studies, including,[69] have explored systematic differences in reported GABA+ estimates between different manufacturers, and their relation to GABA editing efficiency and the contribution of coedited macromolecules to the measured signal. Furthermore, a previous examination of water-scaled GABA+ estimates on a superset of data also incorporating the present study's subjects[24] identified systematically higher GABA+ estimates from datasets acquired on the Siemens platform, by approximately 29%, which could not be explained in terms of editing efficiency or macromolecule contribution. Observations in the present analysis corroborate this, with most algorithms yielding higher GABA+/H_2O estimates from Siemens datasets, and all algorithms yielding higher estimates for both Glx_{diff}/H_2O and tCr_{edit_off}/H_2O for Siemens datasets. The fact that this trend is seen across different metabolites and within the edit-OFF subspectra gives support to the notion that water reference data from the Siemens implementation may not be optimal for scaling purposes, although it cannot be ruled out that both GE and Philips share a similar mis-scaling.

4.3 | Macromolecule fitting

The observed signal around 3 ppm includes substantial contamination (around 50%[38–42]) from coedited signals, including homocarnosine,[70] with coupling between 1.89 and 3.00 ppm spins, and poorly characterised components tentatively attributed to lysine-containing macromolecules,[6,39,71,72] with coupling between 1.71 and 3.01 ppm spins. This may give rise to large residuals or biased baseline estimates if not considered in the model. Although their impact has been studied by some,[39,41,43–45,73] there is currently no consensus on how these should be handled.

The default LCModel configuration (flat baseline) performs surprisingly well by simply ignoring such signals, giving rise to characteristic peaks in the residuals. Meanwhile, the case of Osprey (without MM3co component, Figure 2F) exemplifies the potential baseline distortions and correspondingly reduced GABA+ estimates resulting from these residuals. The result of our ICC analysis across all algorithms suggests that more consistent GABA+ estimates may be obtained by explicitly parametrizing the MM3 contribution in the model. However, using correlation between GABA+ estimates and grey matter fraction as a benchmark, incorporation of the MM3 component did not significantly impact the effectiveness of individual algorithms in measuring differences in GABA+ levels. Moreover, while the supplementary analysis (supporting information, section B) suggested improved effectiveness for LCModel after incorporation of soft constraints on MM3 amplitude (whether to GABA or MM0.9), similar performance for this algorithm was obtained by simply modelling a stiff but nonzero baseline, allowing this to absorb some of the MM3 signal. This configuration was also found to be effective for modelling GABA alone, consistent with previously reported findings,[73] wherein modelling a more flexible baseline in LCModel to selectively remove a portion of the MM3 contribution allowed for closer measurement of GABA

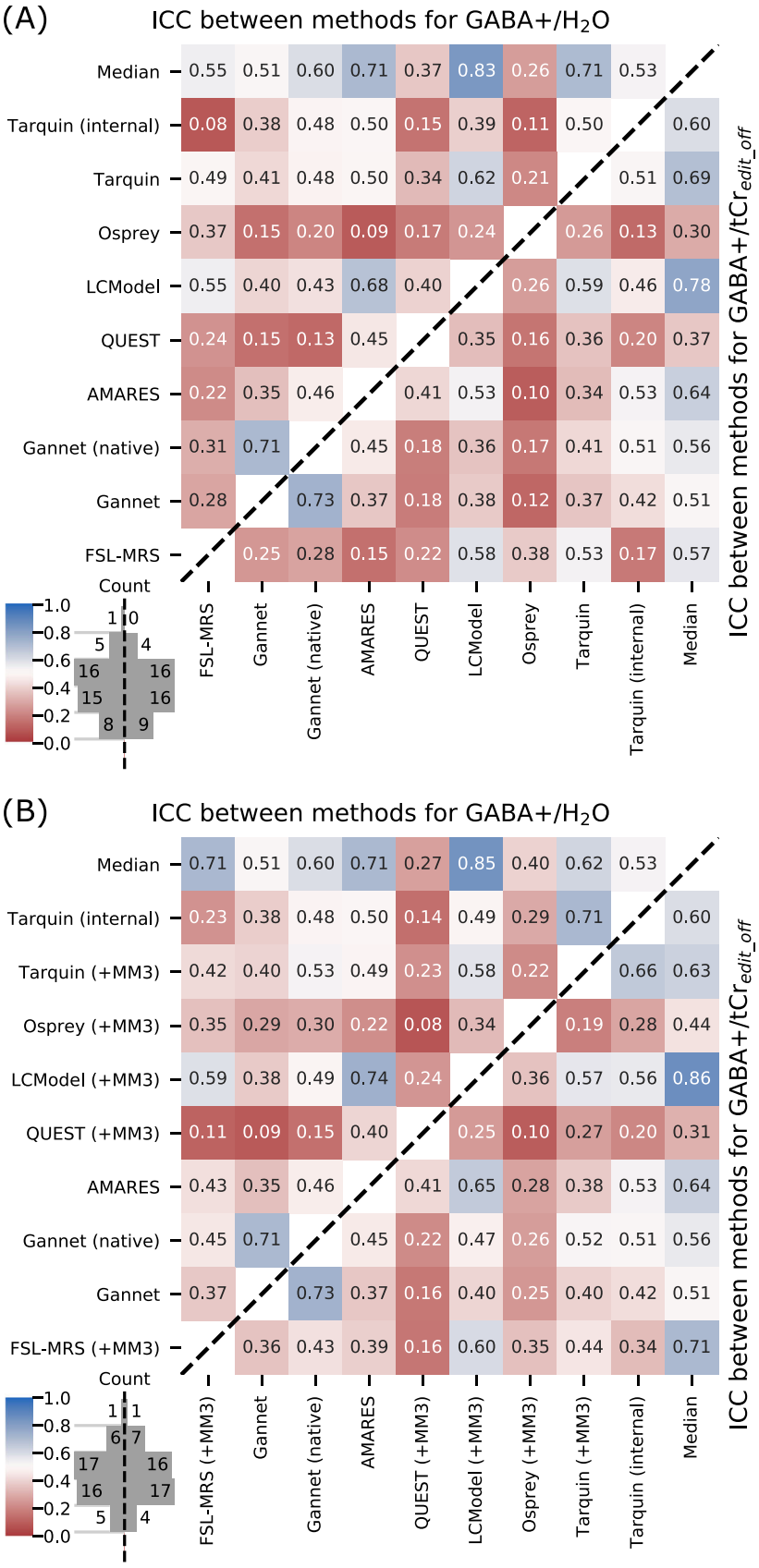

FIGURE 5 Intraclass correlation coefficients between algorithms, scaled to water (upper left triangle) and tCr_{edit_off} (lower right triangle), with basis set algorithms (A) Excluding and (B) Including a component representing the coedited macromolecule contribution. 'Median' data denote correlation with the median estimate across all algorithms

rather than GABA+. Estimates from Osprey in similar configurations were comparable, and it has been shown that higher degrees of baseline flexibility cause greater fractions of the 3-ppm signal to be absorbed into the baseline.[43]

Peak-fitting algorithms (such as Gannet and AMARES) may circumvent this issue somewhat by considering the entire 3.0 ppm GABA+MM signal with a single broad Gaussian model as examined herein; this approach performed comparably with more elaborate models. Furthermore, Tarquin with its internally simulated basis set has two separate Gaussian components representing the GABA+ signal, which are seen to shift and broaden to conform to the shape of the observed GABA+MM signal (Figure 2H). QUEST, similarly, broadened the GABA basis function substantially to more closely envelop the entire GABA+MM signal and (unfortunately) adjacent artefacts (Figure 2D), perhaps accounting for its somewhat lower correlation with grey matter fraction and agreement with other algorithms.

4.4 | Artefact rejection around 3.2 ppm

MEGA-edited GABA spectra often exhibit a slight artefactual feature around 3.2 ppm, which can be problematic for fitting algorithms. The origin is uncertain, but potentially related to incomplete subtraction of Cho[74] or contribution from undetermined other coedited signals (such as, perhaps, valine-containing macromolecules[71] or arginine[75]). In the present study, the baseline for the Osprey fit (without MM3co) tends to respond to this artefact, inducing a bend in the baseline that appears to cut out a significant part of the real GABA peak (Figure 2F), leading to a likely underestimation of GABA+ area. QUEST appears to broaden the GABA and/or MM3co basis components, incorporating the artefact into the GABA+ estimate and most likely overestimating the GABA+ signal area (Figure 2D); this effect was most pronounced for Siemens datasets (see Figure S9D), where the feature manifests more prominently. FSL-MRS and LCModel both handle the artefact well in the general case, largely rejecting it from both the baseline and metabolite models (Figure 2A,E); this is likely attributable to the fact that their default MEGA-PRESS settings prescribe a low-order polynomial baseline (FSL-MRS) or no baseline at all (LCModel). While other basis set algorithms end up somewhere in between (with a degree of contamination from the artefact), peak-fitting algorithms AMARES and Gannet both perform well in this area. Indeed, Gannet (Figure 2B) is the only algorithm to explicitly deal with this artefact, down-weighting some residuals in this region. We suggest that comparably rigid baseline estimation as well as incorporating a Gaussian basis component around 3.2 ppm, with tight constraints on linewidth, shift and amplitude to avoid inadvertently fitting part of the GABA peak, may yield some benefits in this area for other algorithms. Ultimately, further investigation into the underlying signal, and more complete profiling of the coedited metabolite and macromolecule signals in the region, would be preferable.

4.5 | Glx

Although quantification of Glx from the difference spectrum has been demonstrated to be reliable given suitable quality constraints,[76] several researchers have highlighted the relatively low concordance between estimates from short-TE PRESS spectra and GABA-edited difference spectra, with estimates from the edit-OFF subspectra often found to agree better with the short-TE PRESS[77–79]; this is unsurprising given that the Glx signal in the short-TE and edit-OFF subspectra are subject to similar underlying uncertainties, including MM background and overlapping GSH and aspartyl signals. In comparing Glx quantification between MEGA-PRESS difference and edit-OFF subspectra, recent studies[79,80] report a correlation around $r = 0.8$; results in the present study show a more moderate correlation, between $r = 0.34$ and 0.69 depending on algorithm (see Figure S13); a linear scaling factor is also observed, consistent with recent findings.[81] It is notable that agreement between algorithms is higher for coedited Glx than for GABA+, reflecting the better-defined signal seen in the difference spectrum (Figure 2).

With reference to Figure S8, all basis set algorithms showed a distinct structure in the residuals around 3.7 ppm, with the model peaks appearing a little to the right of the peaks observed in the data. This is most likely due to the complicated signal patterns around 2.3 ppm in the edited spectrum (resulting from overlapping signals of GABA and coedited Glu, Gln and GSH), which interact critically with the 3.75 ppm modelling. It is likely that there is a poorly understood baseline fluctuation arising from coedited macromolecular signals appearing between 1.5 and 2.5 ppm,[45] which will bias the correct phase estimation of the 2.25 ppm signals, at the expense of getting the phase of the related 3.75 ppm signals right. The peak-fitting algorithms tested, where modelling around 3.7 ppm is not bound to features in other parts of the spectrum (such as around 2.3 ppm), show much lower residuals in the region. It is possible that basis set fitting on a constrained range would mitigate this effect, at the expense of throwing away useful spectral information and hence detracting from the utility of the basis set approach in general. A model that shares lineshape information between the 2.3 and 3.7 ppm Glx peaks but allows a tightly constrained frequency shift between them may present a reasonable alternative.

4.6 | Limitations

The basis set adopted in the present study was simulated with ideal excitation pulses, and therefore may not fully model subtle variations in spectral structure between manufacturers. However, the impact of excitation is likely negligible compared with the impact of refocusing, which is

appropriately accounted for in the 2D simulations. Vendor-specific excitation pulses may also contribute to the subtly different shape and asymmetry of the 3.0 ppm peak and varying manifestation of the 3.2 ppm feature, which may be observed in Figure S9.

While the present analysis examines a variety of commonly used implementations representing a range of modelling strategies for GABA-edited spectroscopy data, we note that several other algorithms and implementations are also available to the MRS community, including AQSES,[82] INSPECTOR,[83] KALPANA,[84] OXSA,[85] spant[86] and Vespa.[87]

Furthermore, many of the packages examined offer extended functionality, which may well lead to improved performance in certain circumstances, but this did not align with our approach of adopting recommended/default configurations for all algorithms. Most significantly, many software packages offer the fine-tuning of several aspects of the modelling process, for instance, the baseline parametrization. As further examples, jMRUI QUEST offers flexible baseline modelling strategies; FSL-MRS offers independent shift groups that were not assessed; and Osprey can additionally simultaneously optimise difference and sum spectra, potentially benefitting from additional spectral information and improved SNR. Both peak-fitting algorithms examined are flexible in their choice of model, with, for example, dual-Gaussian models for the GABA+ signal readily available. An inevitable consequence of adopting default settings in this analysis is that these might not be optimal for cross-vendor data processed with the standard pipeline adopted herein. All the tools tested are highly configurable and offer expert users many possibilities to tune performance optimally for particular datasets, offering the potential for further invention and protection against establishing a possibly incorrect orthodoxy. Nonetheless, this flexibility comes with the caveat that it may also lead to misuse, selective reporting and inappropriate modelling. Moreover, such variability runs counter to efforts to standardise analysis methodology.[1] While more research into optimised modelling strategies is needed to improve the comparability and robustness of MEGA-PRESS fitting, this work highlights that the complete and accurate reporting of all decisions made during analysis and modelling is immensely important, going even beyond the recently published minimum reporting standards for MRS.[88]

The present study examines metabolite estimates for each algorithm, relative to water and Cr references obtained using that same algorithm: this reflects typical usage, but means that variations discussed herein are not necessarily driven purely by the metabolite modelling.

Finally, because the findings documented herein are substantiated entirely by in vivo data, there is no 'ground truth' available by which to directly assess the algorithms' accuracy. While this limitation has been partially addressed by considering the strength of correlation between GABA+ estimates and grey matter fraction as an index of relative accuracy,[35] a more direct assessment of accuracy could be achieved in further studies involving carefully prepared phantom or synthetic data,[89] each approach having its own inherent limitations. In either case, meticulous attention to macromolecule baseline, SNR and line shape would be required to ensure transferability of findings to in vivo applications.

4.7 | Key recommendations

Based on these findings, we recommend the following for future studies:

- When applying basis set modelling approaches, special consideration must be given to the coedited macromolecular signal underlying the 3.0 ppm GABA peak. While appropriate modelling outcomes may be obtained in some cases by entrusting a carefully tuned baseline to capture the entirety of the signal,[43] a more generalisable approach is simply to routinely incorporate a simulated basis component to represent this signal.
- Care must be taken to ensure consistent behaviour in the presence of commonly observed artefacts, such as the signal around 3.2 ppm. This artefact could be explicitly incorporated into the model, or mitigated with a rigid baseline model, which is less likely to follow the local signal curvature.
- More generally, when inspecting fit outcomes, the behaviour of the baseline (where modelled) demands close attention, to ensure that it does not unduly bias modelling of the GABA peak.
- When assessing data acquired at different sites, systematic differences in scale are to be expected and must be considered, regardless of the algorithm applied.

Additionally, we propose four key areas for further systematic investigation in future studies:

1. Robust methods for the generation of large synthetic datasets for validation are necessary to facilitate direct assessment of modelling accuracy. These synthetic data need to be truly representative of in vivo data, hence the design of their underlying physical signal models will require great attention to detail. Subsequent interpretation must bear in mind that the outcome is likely to be determined by the degree of similarity between the physical model used to generate the data and the model used to decompose it during linear-combination fitting.
2. Detailed exploration of the coedited macromolecule profile that underlies typical GABA-edited data is required. Metabolite-nulled edited spectra obtained on different hardware platforms could provide the basis for a more informed parametrization of these signals during modelling, and also contribute to accurate and in vivo-like representation of synthetic data.

3. The origin of the spectral feature around 3.2 ppm requires further investigation: it needs to be determined whether this is a subtraction artefact or an actual real signal (e.g., from valine-containing macromolecules or other signals hitherto not routinely included in modelling, such as arginine, which has a compatible spin system). Clarification of this feature would allow for more appropriate modelling and parametrization of synthetic data.
4. When considering Glx estimates from the difference spectrum, further investigation into the complex interactions of coedited Glu, Gln, GSH and possibly other signals around 2.3 ppm, and their impact on the 3.75 ppm Glx modelling, would be beneficial. This focus area will benefit from increased insight into the macromolecular background of the GABA-edited spectrum.

4.8 | Conclusions

Although the observed consistency across algorithms was generally moderate, with pairwise correlation in some cases concerningly weak, we emphasise that more consistent estimates are not necessarily more accurate estimates: all the algorithms tested (except for QUEST without MM3) were shown to be comparable in their effectiveness in detecting differences in GABA+ concentration. This does, however, raise some concerns regarding the comparability of findings between different studies, each of which will typically employ a single modelling algorithm, often with divergent processing and prior knowledge and with significantly smaller sample sizes than tested here. This means that the choice of analysis already adds considerable uncertainty and variability.

Improved standardisation of sequence implementation,[27,90] and adoption of standardised processing pipelines and prior knowledge (e.g., in the form of publicly available basis set libraries), may reduce sources of variation between studies. However, the interaction of baseline, coedited macromolecule and metabolite signals, and other artefactual signals, remains a critical source of variation between algorithms, and within different configurations of the same algorithm. Better characterisation of these signals would allow for more complete modelling (at the risk of overfitting). Consensus on optimal (or at least, appropriate) control parameters for the respective algorithms would also be beneficial; this should be informed by representative datasets covering multiple sites and vendors, as facilitated by publicly available repositories, such as the one leveraged herein.[24,25] It may further be of benefit to conceive a 'consensus algorithm' to be implemented across software environments, and used as a shared starting point to refine the algorithmic decision-making in future iterations of the algorithm. Meanwhile, careful attention to the behaviour of the model with regard to such signals, and rigorous reporting of the configuration employed, are necessary to facilitate meaningful comparisons between studies.

ACKNOWLEDGMENTS

The authors wish to thank Dr Stephen Provencher for his assistance and constructive feedback on the application of the LCModel algorithm. The graphical abstract was illustrated by Laura Garrison (University of Bergen). Analysis was performed within a project funded by the H2020 European Research Council Advanced Grant #693124, which additionally supports the contributions of ARC, LE and KH. Data used in this analysis were previously collected through an international collaborative study funded under National Institutes of Health (NIH) grant R01 EB016089. WTC is supported by funding from Wellcome Trust and the Royal Society (102584/Z/13/Z). PM thanks India–Australia Strategic Biotechnology Funding (BT/Indo-Aus/10/31/2016/PKM). RR was supported by the Australian Research Council (DE210100790). GO and RE acknowledge funding support from NIH grants K99/R00 AG062230, S10 OD021648, P41 EB031771, P41 EB015909, R01 EB016089, R01 EB023963, R01 EB028259 and R21 HD100869.

CONFLICT OF INTEREST

The authors declare no conflicting interests.

DATA AVAILABILITY STATEMENT

Scripts used for automation and reporting contained in the present manuscript are publicly available here; further dependencies are described within: https://git.app.uib.no/bergen-fmri/analyzing-big-gaba. Spectra analysed in this manuscript were obtained from the publicly available Big GABA repository on NITRC, https://www.nitrc.org/projects/biggaba. Basis sets used in the primary analysis were obtained from the publicly available Osprey package, https://schorschinho.github.io/osprey.

ORCID

Alexander R. Craven https://orcid.org/0000-0003-2583-7571
William T. Clarke https://orcid.org/0000-0001-7159-7025
Richard A. E. Edden https://orcid.org/0000-0002-0671-7374
Pravat K. Mandal https://orcid.org/0000-0003-4999-2808
Mark Mikkelsen https://orcid.org/0000-0002-0349-3782

James B. Murdoch https://orcid.org/0000-0001-7303-1914
Jamie Near https://orcid.org/0000-0003-3516-936X
Reuben Rideaux https://orcid.org/0000-0001-8416-005X
Martin Wilson https://orcid.org/0000-0002-2089-3956
Helge J. Zöllner https://orcid.org/0000-0002-7148-292X
Georg Oeltzschner https://orcid.org/0000-0003-3083-9811

REFERENCES

1. Bhogal AA, Schür RR, Houtepen LC, et al. ^{1}H-MRS processing parameters affect metabolite quantification: The urgent need for uniform and transparent standardization. *NMR Biomed.* 2017;30(11):e3804. doi:10.1002/nbm.3804
2. Kanowski M, Kaufmann J, Braun J, Bernarding J, Tempelmann C. Quantitation of simulated short echo time1H human brain spectra by LCModel and AMARES. *Magn Reson Med.* 2004;51(5):904-912. doi:10.1002/mrm.20063
3. Mullins PG, Rowland L, Bustillo J, Bedrick EJ, Lauriello J, Brooks WM. Reproducibility of 1H-MRS measurements in schizophrenic patients. *Magn Reson Med.* 2003;50(4):704-707. doi:10.1002/mrm.10598
4. Zöllner HJ, Považan M, Hui SCN, Tapper S, Edden RAE, Oeltzschner G. Comparison of different linear-combination modeling algorithms for short-TE proton spectra. *NMR Biomed.* 2021;34(4):e4482. doi:10.1002/nbm.4482
5. Mescher M, Merkle H, Kirsch J, Garwood M, Gruetter R. Simultaneous in vivo spectral editing and water suppression. *NMR Biomed.* 1998;11(6):266-272. doi:10.1002/(sici)1099-1492(199810)11:6%3C266::aid-nbm530%3E3.0.co;2-j
6. Rothman DL, Petroff OA, Behar KL, Mattson RH. Localized 1H NMR measurements of gamma-aminobutyric acid in human brain in vivo. *Proc Natl Acad Sci.* 1993;90(12):5662-5666. doi:10.1073/pnas.90.12.5662
7. Brix MK, Ersland L, Hugdahl K, et al. Within- and between-session reproducibility of GABA measurements with MR spectroscopy: Reproducibility of MRS GABA measurements. *J Magn Reson Imaging.* 2017;46(2):421-430. doi:10.1002/jmri.25588
8. Duda JM, Moser AD, Zuo CS, et al. Repeatability and reliability of GABA measurements with magnetic resonance spectroscopy in healthy young adults. *Magn Reson Med*. 2021;85:2359-2369. doi:10.1002/mrm.28587
9. O'Gorman RL, Michels L, Edden RA, Murdoch JB, Martin E. In vivo detection of GABA and glutamate with MEGA-PRESS: Reproducibility and gender effects. *J Magn Reson Imaging.* 2011;33(5):1262-1267. doi:10.1002/jmri.22520
10. Baeshen A, Wyss PO, Henning A, et al. Test–retest reliability of the brain metabolites GABA and Glx with JPRESS, PRESS, and MEGA-PRESS MRS sequences in vivo at 3T. *J Magn Reson Imaging.* 2020;51(4):1181-1191. doi:10.1002/jmri.26921
11. Zöllner HJ, Oeltzschner G, Schnitzler A, Wittsack HJ. In silico GABA+ MEGA-PRESS: Effects of signal-to-noise ratio and linewidth on modeling the 3 ppm GABA+ resonance. *NMR Biomed*. 2021;34:e4410. doi:10.1002/nbm.4410
12. Mikkelsen M, Bhattacharyya PK, Mandal PK, et al. Analyzing big GABA: Comparison of five software packages for GABA-edited MRS. Paper presented at 27th ISMRM Annual Meeting, Montréal, Canada; 2019.
13. Clarke WT, Stagg CJ, Jbabdi S. FSL-MRS: An end-to-end spectroscopy analysis package. *Magn Reson Med*. 2021;85(6):2950-2964. doi:10.1002/mrm.28630
14. Edden RAE, Puts NAJ, Harris AD, Barker PB, Evans CJ. Gannet: A batch-processing tool for the quantitative analysis of gamma-aminobutyric acid-edited MR spectroscopy spectra: Gannet: GABA Analysis Toolkit. *J Magn Reson Imaging.* 2014;40(6):1445-1452. doi:10.1002/jmri.24478
15. Provencher SW. Estimation of metabolite concentrations from localized in vivo proton NMR spectra. *Magn Reson Med.* 1993;30(6):672-679. doi:10.1002/mrm.1910300604
16. Oeltzschner G, Zöllner HJ, Hui SCN, et al. Osprey: Open-source processing, reconstruction & estimation of magnetic resonance spectroscopy data. *J Neurosci Methods.* 2020;343:108827. doi:10.1016/j.jneumeth.2020.108827
17. Reynolds G, Wilson M, Peet A, Arvanitis TN. An algorithm for the automated quantitation of metabolites in in vitro NMR signals. *Magn Reson Med.* 2006;56(6):1211-1219. doi:10.1002/mrm.21081
18. Wilson M, Reynolds G, Kauppinen RA, Arvanitis TN, Peet AC. A constrained least-squares approach to the automated quantitation of in vivo ^{1}H magnetic resonance spectroscopy data. *Magn Reson Med.* 2011;65(1):1-12. doi:10.1002/mrm.22579
19. Vanhamme L, van den Boogaart A, Van Huffel S. Improved method for accurate and efficient quantification of MRS data with use of prior knowledge. *J Magn Reson.* 1997;129(1):35-43. doi:10.1006/jmre.1997.1244
20. Graveron-Demilly D. Quantification in magnetic resonance spectroscopy based on semi-parametric approaches. *Magn Reson Mater Phys Biol Med.* 2014;27(2):113-130. doi:10.1007/s10334-013-0393-4
21. Ratiney H, Coenradie Y, Cavassila S, van Ormondt D, Graveron-Demilly D. Time-domain quantitation of 1 H short echo-time signals: background accommodation. *MAGMA Magn Reson Mater Phys Biol Med.* 2004;16(6):284-296. doi:10.1007/s10334-004-0037-9
22. Naressi A, Couturier C, Devos JM, et al. JAVA-based graphical user interface for the MRUI quantitation package. *Magma Magn Reson Mater Phys Biol Med.* 2001;12(2–3):141-152. doi:10.1007/BF02668096
23. Stefan D, Cesare FD, Andrasescu A, et al. Quantitation of magnetic resonance spectroscopy signals: the jMRUI software package. *Meas Sci Technol.* 2009;20(10):104035. doi:10.1088/0957-0233/20/10/104035
24. Mikkelsen M, Rimbault DL, Barker PB, et al. Big GABA II: Water-referenced edited MR spectroscopy at 25 research sites. *Neuroimage.* 2019;191:537-548. doi:10.1016/j.neuroimage.2019.02.059
25. Mikkelsen M, Barker PB, Bhattacharyya PK, et al. Big GABA: edited MR spectroscopy at 24 research sites. *Neuroimage.* 2017;159:32-45. doi:10.1016/j.neuroimage.2017.07.021
26. Považan M, Mikkelsen M, Berrington A, et al. Comparison of multivendor single-voxel MR spectroscopy data acquired in healthy brain at 26 sites. *Radiology*. 2020;295(1):171-180. doi:10.1148/radiol.2020191037
27. Saleh MG, Rimbault D, Mikkelsen M, et al. Multi-vendor standardized sequence for edited magnetic resonance spectroscopy. *Neuroimage.* 2019;189:425-431. doi:10.1016/j.neuroimage.2019.01.056

28. Near J, Harris AD, Juchem C, et al. Preprocessing, analysis and quantification in single-voxel magnetic resonance spectroscopy: experts' consensus recommendations. *NMR Biomed.* 2021;34(5):e4257. doi:10.1002/nbm.4257
29. Simpson R, Devenyi GA, Jezzard P, Hennessy TJ, Near J. Advanced processing and simulation of MRS data using the FID appliance (FID-A)—an open source, MATLAB-based toolkit. *Magn Reson Med.* 2017;77(1):23-33. doi:10.1002/mrm.26091
30. Near J, Edden R, Evans CJ, Paquin R, Harris A, Jezzard P. Frequency and phase drift correction of magnetic resonance spectroscopy data by spectral registration in the time domain: MRS Drift Correction Using Spectral Registration. *Magn Reson Med.* 2015;73(1):44-50. doi:10.1002/mrm.25094
31. Klose U. In vivo proton spectroscopy in presence of eddy currents. *Magn Reson Med.* 1990;14(1):26-30. doi:10.1002/mrm.1910140104
32. Barkhuijsen H, de Beer R, van Ormondt D. Improved algorithm for noniterative time-domain model fitting to exponentially damped magnetic resonance signals. *J Magn Reson.* 1987;73(3):553-557. doi:10.1016/0022-2364(87)90023-0
33. Kreis R. Issues of spectral quality in clinical1H-magnetic resonance spectroscopy and a gallery of artifacts. *NMR Biomed.* 2004;17(6):361-381. doi:10.1002/nbm.891
34. Gasparovic C, Song T, Devier D, et al. Use of tissue water as a concentration reference for proton spectroscopic imaging. *Magn Reson Med.* 2006;55(6):1219-1226. doi:10.1002/mrm.20901
35. Rideaux R, Mikkelsen M, Edden RAE. Comparison of methods for spectral alignment and signal modelling of GABA-edited MR spectroscopy data. *Neuroimage.* 2021;232:117900. doi:10.1016/j.neuroimage.2021.117900
36. Zhang Y, An L, Shen J. Fast computation of full density matrix of multispin systems for spatially localized *in vivo* magnetic resonance spectroscopy. *Med Phys.* 2017;44(8):4169-4178. doi:10.1002/mp.12375
37. Kaiser LG, Young K, Meyerhoff DJ, Mueller SG, Matson GB. A detailed analysis of localized J-difference GABA editing: theoretical and experimental study at 4 T. *NMR Biomed.* 2008;21(1):22-32. doi:10.1002/nbm.1150
38. Choi I, Andronesi OC, Barker P, et al. Spectral editing in ^{1}H magnetic resonance spectroscopy: Experts' consensus recommendations. *NMR Biomed.* 2020;18:e4411. doi:10.1002/nbm.4411
39. Deelchand DK, Marjańska M, Henry P, Terpstra M. MEGA-PRESS of GABA+: Influences of acquisition parameters. *NMR Biomed.* 2021;34(5):e4199. doi:10.1002/nbm.4199
40. Shungu DC, Mao X, Gonzales R, et al. Brain γ-aminobutyric acid (GABA) detection *in vivo* with the *J* -editing ^{1}H MRS technique: a comprehensive methodological evaluation of sensitivity enhancement, macromolecule contamination and test-retest reliability. *NMR Biomed.* 2016;29(7):932-942. doi:10.1002/nbm.3539
41. Henry PG, Dautry C, Hantraye P, Bloch G. Brain GABA editing without macromolecule contamination. *Magn Reson Med.* 2001;45(3):517-520. doi:10.1002/1522-2594(200103)45:3%3C517::aid-mrm1068%3E3.0.co;2-6
42. Mullins PG, McGonigle DJ, O'Gorman RL, et al. Current practice in the use of MEGA-PRESS spectroscopy for the detection of GABA. *Neuroimage.* 2014;86:43-52. doi:10.1016/j.neuroimage.2012.12.004
43. Zöllner HJ, Tapper S, Hui SCN, Barker PB, Edden RAE, Oeltzschner G. Comparison of linear combination modeling strategies for edited magnetic resonance spectroscopy at 3 T. *NMR Biomed.* 2022;35(1):e4618. doi:10.1002/nbm.4618
44. Murdoch JB, Dydak U. Modeling MEGA-PRESS macromolecules for a better grasp of GABA. Proc Int Soc Magn Reson Med. 2011;19:1394. https://cds.ismrm.org/protected/11MProceedings/files/1394.pdf
45. Bhagwagar Z, Wylezinska M, Jezzard P, et al. Reduction in occipital cortex γ-aminobutyric acid concentrations in medication-free recovered unipolar depressed and bipolar subjects. *Biol Psychiatry.* 2007;61(6):806-812. doi:10.1016/j.biopsych.2006.08.048
46. Pedrosa de Barros N, Slotboom J. Quality management in in vivo proton MRS. *Anal Biochem.* 2017;529:98-116. doi:10.1016/j.ab.2017.01.017
47. Kreis R. The trouble with quality filtering based on relative Cramér-Rao lower bounds. *Magn Reson Med.* 2016;75(1):15-18. doi:10.1002/mrm.25568
48. Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. *WIREs Data Min Knowl Discov.* 2011;1(1):73-79. doi:10.1002/widm.2
49. McKinney W. Data structures for statistical computing in Python. Proceedings, Python in Science Conference, Austin, Texas. 2010;56–61. doi:10.25080/Majora-92bf1922-00a
50. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature.* 2020;585(7825):357-362. doi:10.1038/s41586-020-2649-2
51. SciPy 1.0 Contributors, Virtanen P, Gommers R, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17(3):261-272. doi:10.1038/s41592-019-0686-2
52. Vallat R. Pingouin: statistics in Python. *J Open Source Softw.* 2018;3(31):1026. doi:10.21105/joss.01026
53. Seabold S, Perktold J. statsmodels: Econometric and statistical modeling with python; 9th Python in Science Conference; 2010. https://www.statsmodels.org/. Accessed 13 May 2021.
54. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika.* 1965;52(3/4):591. doi:10.2307/2333709
55. Fligner MA, Killeen TJ. Distribution-free two-sample tests for scale. *J Am Stat Assoc.* 1976;71(353):210-213. doi:10.1080/01621459.1976.10481517
56. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet Lond Engl.* 1986;1(8476):307-310.
57. Welch BL. The generalisation of student's problems when several different population variances are involved. *Biometrika.* 1947;34(1–2):28-35. doi:10.1093/biomet/34.1-2.28
58. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat.* 1979;6(2):65-70.
59. Bonferroni CE. Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del Professore Salvatore Ortu Carboni.* Rome: Tip. del Senato; 1935;13-60.
60. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing; 2018. https://www.R-project.org/. Accessed 30 July 2021.
61. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw.* 2015;67(1):1-48. doi:10.18637/jss.v067.i01
62. Halekoh U, Højsgaard S. A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models - the *R* package pbkrtest. *J Stat Softw.* 2014;59(9):1-32. doi:10.18637/jss.v059.i09
63. Harris AD, Puts NAJ, Edden RAE. Tissue correction for GABA-edited MRS: Considerations of voxel composition, tissue segmentation, and tissue relaxations. *J Magn Reson Imaging.* 2015;42(5):1431-1440. doi:10.1002/jmri.24903
64. Jensen JE, de Frederick B, Renshaw PF. Grey and white matter GABA level differences in the human brain using two-dimensional, J-resolved spectroscopic imaging. *NMR Biomed.* 2005;18(8):570-576. doi:10.1002/nbm.994

65. Mikkelsen M, Singh KD, Brealy JA, Linden DEJ, Evans CJ. Quantification of γ-aminobutyric acid (GABA) in ^{1}H MRS volumes composed heterogeneously of grey and white matter. *NMR Biomed.* 2016;29(11):1644-1655. doi:10.1002/nbm.3622
66. Pernet CR, Wilcox R, Rousselet GA. Robust correlation analyses: false positive and power validation using a new open source Matlab toolbox. *Front Psychol.* 2013;3. doi:10.3389/fpsyg.2012.00606
67. Rousselet GA, Pernet CR. Improving standards in brain-behavior correlation analyses. *Front Hum Neurosci.* 2012;6. doi:10.3389/fnhum.2012.00119
68. Landheer K, Juchem C. Are Cramér-Rao lower bounds an accurate estimate for standard deviations in in vivo magnetic resonance spectroscopy? *NMR Biomed.* 2021;19:e4521. doi:10.1002/nbm.4521
69. Harris AD, Puts NA, Wijtenburg SA, et al. Normalizing data from GABA-edited MEGA-PRESS implementations at 3 Tesla. *Magn Reson Imaging.* 2017;42:8-15. doi:10.1016/j.mri.2017.04.013
70. Rothman DL, Behar KL, Prichard JW, Petroff OAC. Homocarnosine and the measurement of neuronal pH in patients with epilepsy. *Magn Reson Med.* 1997;38(6):924-929. doi:10.1002/mrm.1910380611
71. Behar KL, Ogino T. Characterization of macromolecule resonances in the ^{1}H NMR spectrum of rat brain. *Magn Reson Med.* 1993;30(1):38-44. doi:10.1002/mrm.1910300107
72. Cudalbu C, Behar KL, Bhattacharyya PK, et al. Contribution of macromolecules to brain ^{1} H MR spectra: Experts' consensus recommendations. *NMR Biomed.* 2020;25:e4393. doi:10.1002/nbm.4393
73. Dydak U, Jiang YM, Long LL, et al. In vivo measurement of brain GABA concentrations by magnetic resonance spectroscopy in smelters occupationally exposed to manganese. *Environ Health Perspect.* 2011;119(2):219-224. doi:10.1289/ehp.1002192
74. Evans CJ, Puts NAJ, Robson SE, et al. Subtraction artifacts and frequency (mis-)alignment in *J*-difference GABA editing. *J Magn Reson Imaging.* 2013;38(4):970-975. doi:10.1002/jmri.23923
75. Jofre F, Anderson ME, Markley JL. L_arginine. Biological Magnetic Resonance Bank; 2006. doi:10.13018/BMSE000029
76. Sanaei Nezhad F, Anton A, Michou E, Jung J, Parkes LM, Williams SR. Quantification of GABA, glutamate and glutamine in a single measurement at 3 T using GABA-edited MEGA-PRESS. *NMR Biomed.* 2018;31(1):e3847. doi:10.1002/nbm.3847
77. Bell T, Boudes ES, Loo RS, et al. In vivo Glx and Glu measurements from GABA-edited MRS at 3 T. *NMR Biomed.* 2020;e4245. doi:10.1002/nbm.4245
78. Maddock RJ, Caton MD, Ragland JD. Estimating glutamate and Glx from GABA-optimized MEGA-PRESS: Off-resonance but not difference spectra values correspond to PRESS values. *Psychiatry Res Neuroimaging.* 2018;279:22-30. doi:10.1016/j.pscychresns.2018.07.003
79. van Veenendaal TM, Backes WH, van Bussel FCG, et al. Glutamate quantification by PRESS or MEGA-PRESS: Validation, repeatability, and concordance. *Magn Reson Imaging.* 2018;48:107-114. doi:10.1016/j.mri.2017.12.029
80. Dhamala E, Abdelkefi I, Nguyen M, Hennessy TJ, Nadeau H, Near J. Validation of *in vivo* MRS measures of metabolite concentrations in the human brain. *NMR Biomed.* 2019;32(3):e4058. doi:10.1002/nbm.4058
81. Cheng H, Wang A, Newman S, Dydak U. An investigation of glutamate quantification with PRESS and MEGA-PRESS. *NMR Biomed.* 2021;34(2):e4453. doi:10.1002/nbm.4453
82. Poullet JB, Sima DM, Simonetti AW, et al. An automated quantitation of short echo time MRS spectra in an open source software environment: AQSES. *NMR Biomed.* 2007;20(5):493-504. doi:10.1002/nbm.1112
83. Gajdošík M, Landheer K, Swanberg KM, Juchem C. INSPECTOR: free software for magnetic resonance spectroscopy data inspection, processing, simulation and analysis. *Sci Rep.* 2021;11(1):2094. doi:10.1038/s41598-021-81193-9
84. Mandal PK, Shukla D. KALPANA: advanced spectroscopic signal processing platform for improved accuracy to aid in early diagnosis of brain disorders in clinical setting. *J Alzheimers Dis.* 2020;75(2):397-402. doi:10.3233/JAD-191351
85. Purvis LAB, Clarke WT, Biasiolli L, Valkovič L, Robson MD, Rodgers CT. OXSA: An open-source magnetic resonance spectroscopy analysis toolbox in MATLAB. *PLOS One.* 2017;12(9):e0185356. doi:10.1371/journal.pone.0185356
86. Wilson M. Adaptive baseline fitting for MR spectroscopy analysis. *Magn Reson Med.* 2021;85(1):13-29. doi:10.1002/mrm.28385
87. Soher B, Semanchuk P, Todd D, Steinberg J, Young K. VeSPA: integrated applications for RF pulse design, spectral simulation and MRS data analysis. *Proc Int Soc Magn Reson Med.* 2011;19:1410-1410.
88. Lin A, Andronesi O, Bogner W, et al. Minimum Reporting Standards for in vivo Magnetic Resonance Spectroscopy (MRSinMRS): Experts' consensus recommendations. *NMR Biomed.* 2021;9:e4484. doi:10.1002/nbm.4484
89. Henning A. Advanced Spectral Quantification: Parameter Handling, Nonparametric Pattern Modeling, and Multidimensional Fitting. In: Harris RK, Wasylishen RL, eds. *EMagRes.* John Wiley & Sons, Ltd; 2016:981-994. doi:10.1002/9780470034590.emrstm1472
90. Deelchand DK, Berrington A, Noeske R, et al. Across-vendor standardization of semi-LASER for single-voxel MRS at 3T. *NMR Biomed.* 2021;34(5):e4218. doi:10.1002/nbm.4218

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Craven AR, Bhattacharyya PK, Clarke WT, et al. Comparison of seven modelling algorithms for γ-aminobutyric acid–edited proton magnetic resonance spectroscopy. *NMR in Biomedicine*. 2022;e4702. doi:10.1002/nbm.4702