Improving reproducibility of proton MRS brain thermometry: theoretical and empirical approaches

Zhengchao Dong[1,2], Joshua T. Kantrowitz[1,2,3], J. John Mann[1,2,4]

1. Department of Psychiatry
   Columbia University College of Physicians & Surgeons, New York, USA
2. New York State Psychiatric Institute, New York, USA
3. Nathan Kline Institute, 140 Old Orangeburg Road, Orangeburg, NY, USA
4. Department of Radiology, Columbia University, College of Physicians and Surgeons, New York, NY, USA.

Correspondence to:

Zhengchao Dong, Ph.D.

Department of Psychiatry

Columbia University

New York, USA

Email: Zhengchao.Dong@nyspi.columbia.edu

Tel. (646) 774-5828

Abstract

Purpose: In ^{1}H MRS-based thermometry of brain, averaging temperatures measured from more than one reference peak offers several advantages including improving the reproducibility, i.e. precision, of the measurement. This paper proposes theoretically and empirically optimal weighting factors to improve the weighted average of temperatures measured from three references.

Methods: We first proposed concepts of equivalent noise and equivalent signal-to-noise ratio in terms of frequency measurement and a concept of relative frequency that allows the combination of different peaks in a spectrum for improving the accuracy of frequency measurement. Based on these, we then developed a theoretically optimal weighting factor and suggested an empirical weighting factor for weighted average of temperatures measured from three references in ^{1}H MRS-based thermometry. We assessed the two new weighting factors, together with other two previously proposed weighting factors, by comparing the errors of temperatures measured from individual references and the errors of averaged temperatures using these differing weighting factors. These errors were defined as the standard deviations in repeated measurements and in Monte Carlo studies. We also performed computer simulations to aid error analyses in temperature averaging.

Results: Both the proposed theoretical and empirical weighting factors outperformed the other two previously proposed weighting factors as well as the three individual references in all phantom and in vivo experiments. In phantom experiments with 4 Hz or 10 Hz line broadening, the theoretical weighting outperformed the empirical one, but the latter was superior in all other repeated and Monte Carlo tests performed on phantom and in vivo data. Computer simulations offered explanations for the performances of the two new proposed weightings.

Conclusion: The proposed two new weighting factors are superior to the two previously proposed weighting factors and can improve the measurement of temperature using ^{1}H MRS-based thermometry.

Key words:

^{1}H MR spectroscopy, thermometry, temperature, frequency, Cramer-Rao lower bound, equivalent noise, optimally weighted average

1. Introduction

The ^{1}H MRS-based thermometry of the brain seeks to measure brain temperature and its distribution. Unlike other MR-based thermometric methods that measure relative temperature changes spatially or temporally, and unlike other invasive techniques that enable measurement of absolute temperature with neurosurgical interventions, the ^{1}H MRS-based thermometry can measure absolute temperature non-invasively[1]. This feature has made this technique especially useful in pre-clinical or clinical physiology and pathophysiological studies such as those measuring temperature of neonatal brain[2], diagnosing intracranial tumors[3], monitoring brain trauma[4], and image-guided thermal ablation[5]. Another less exploited feature of ^{1}H MRS-based thermometry is that it allows quantification of brain metabolites simultaneously, without an additional scan. This permits use of ^{1}H MRS in studying correlations between temperature and metabolism in brain development, treatment and pathophysiology of disorders.[6, 7]

The ^{1}H MRS thermometry measures temperature based on the frequency differences between the temperature-dependent water and temperature-independent references (prominent singlets of metabolites). A commonly used reference is the peak of N-acetyl-aspartate (NAA) at 2.01 ppm. But simultaneous use of multiple references is preferrable because: (1) a preselected reference may not be sufficiently prominent[8] or may even disappear[9]; and (2) weighted average of temperatures derived from multiple references *may* improve reproducibility of temperature measurement.[10] However, the performance of the weighted average of temperatures depends on the averaging algorithms but unoptimized average may be inferior to the best/better single reference.[10]

Here we proposed a concept of *equivalent noise* in terms of frequency measurement based on the Cramer-Rao Lower Bound (CRLB) for frequency and derived an optimized algorithm and presented an empirical algorithm for weighted average of temperatures obtained from multiple references. We assessed our methods and compared their performance with other methods using phantom data and in vivo data from human brain.

2. Methods and materials

2.1 Theory

2.1.1 Cramer-Rao lower bounds and equivalent noise for frequency measurement

We calculated CRLB of frequency measurement for an MRS signal with Lorentzian line shape:

$$s(n) = Ae^{-\alpha n + i\beta n + i\vartheta} + \varepsilon(\mathrm{n}) \qquad [1]$$

$$n = 0, \ldots, N-1$$

where N is the number of points in the time domain signal, $\boldsymbol{A}$, α, β, and ϑ are constants for amplitude, normalized decay, normalized circular frequency and phase, respectively, and ε is the Gaussian noise with standard deviation (SD) σ_0. α and β are related to the linewidth W, resonance frequency f and spectral width sw by the following equations,

$$\alpha = \pi W / sw$$
$$\beta = 2\pi f / sw$$

The CRLB for circular frequency is:

$$CRLB(\beta) = \frac{2\sigma_0\alpha\sqrt{2\alpha}}{A} \quad [2]$$

or

$$CRLB(f) = \sigma_0\alpha \cdot \frac{sw\sqrt{2\alpha}}{A\pi} \quad (error\ in\ Hz\ or\ ppm) \quad [2b]$$

which is the preferred form. The CRLB of frequency is the achievable minimal variance of a measurement, which is proportional to the spectral noise level and inversely proportional to the amplitude of the signal. We defined the CRLB in Eq. 2 as the equivalent noise in terms of frequency measurement:

$$\sigma_e = CRLB(f) = \frac{2\sigma_0\alpha\sqrt{2\alpha}}{A\pi} sw \quad [3]$$

where σ_e is the *equivalent* noise in terms of *frequency* measurement (Figure 1). The above equation indicates that the *equivalent* noises of peaks in the same spectrum may be different if their decay rates and amplitudes are different, although the spectral noise level σ_0 is the same for all. The concept and meaning of the equivalent noise for frequency measurement are better understood from the following example.

2.1.2 Improving frequency measurement by combining/averaging spectra

Signal accumulation or weighted signal averaging is a common practice in NMR for enhancing signal to noise ratio (Figure 1). When we measure the frequencies of the peak in two spectra, the spectrum with higher *SNR* will have smaller error or give higher precision of the measurement; and when we combine the two spectra into one and measure its frequency, we predict, but need to verify, the precision of the measurement will increase if the *SNR* of the combined peak increases.

We now extend this argument to the averaging of multiple peaks in a *single* spectrum by introducing the concept of relative frequencies with respect to true frequencies and employing the concept of equivalent noise (Figure 2). Here the frequencies of different peaks (both their true and measured frequencies$\{f_{i0}\}$ and $\{f_i\}$[11]) are different and the peaks cannot be added. Instead, the relative frequencies $\{f_i - f_{i0}\}$, which are the difference between the measured and the corresponding true frequencies of the individual peaks, are at the same position on the relative frequency axes, subject only to measurement errors, and these peaks are therefore additive, just like the peaks in the above example. The errors of the relative frequency, measured from the combined peaks in the relative axes, can be smaller than those measured from individual peaks.

The conversion of the frequencies into relative frequencies has real, physical meaning in ^{1}H MRST: the temperatures measured from different references are theoretically the same, which are derived from the relative frequency differences of the reference with respect to the same water, subject only to differences stemming from random noise and calibration errors. Therefore, the optimal combination of temperatures measured from references for improving precision is equivalent to optimal combination of peaks on their axes of relative frequencies. The principles of the combinations of the same peak in different acquisitions (Figure 1) and different peaks in the same acquisition (Figure 2) are the same, but the later requires that the noise levels of the

individual reference peaks to be converted to their equivalent noises as outlined in the last subsection. The core focus of the current work is optimal averaging of frequency and eventually temperature measurement. The linear relationship between frequency and temperature, as well as between errors of frequency and temperature measurement, is given in Appendix B. Therefore, the terms of frequency averaging and temperature averaging are interchangeable. Furthermore, the combination is, in essence, averaging in its technical realization, and the terms of "combination of temperatures" and "average of temperatures" are also interchangeable in this paper.

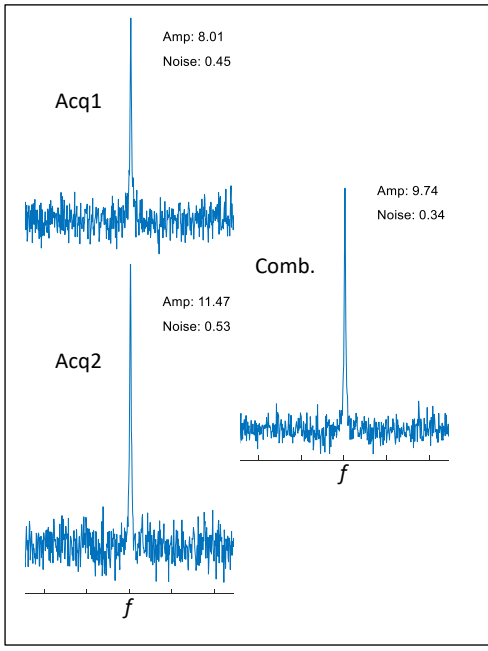

Figure 1. An illustration of signal averaging to enhance the SNR. Shown on the left are two acquisitions from the same sample, with different amplitudes and noise levels. After weighted averaging, the SNR of the combined signal increased (right-hand panel).

2.1.3 Weighted averaging of peaks

The weighted average of the frequencies is expressed as

$$F_c = \sum_{i=1}^{M} k_i F_i \qquad [4a]$$

where F_i is frequency, M is the number of reference peaks, and k_i is the normalized weighting factor:

$$\sum_{1}^{M} k_i = 1 \qquad [4b]$$

Two kinds of weighting factors,

$$k_i = A \text{ and } k_i = A^2 \qquad [5a]$$

were previously proposed.[10] With the concept of equivalent noise, we hypothesize that an optimized weighting factor is one that maximize the equivalent *SNR*, or SNR_e, of the combined peak in terms of frequency measurement.

Suppose we have an MRS signal consisting of M independent, exponentially decaying sinusoids as defined in Eq. 1. For simplicity, we use a 2 x M matrix to describe the amplitudes A and equivalent noises σ_e of the M component signals: (A_1 A_2 … A_M; σ_{e1} σ_{e2} … σ_{eM}). We convert the frequencies of the peaks into the relative frequency axes to facilitate peak combination. Seeking an optimized weighting factor is the key to optimizing the SNR_e of the combined peak. The SNR of the weighted average of the peaks is expressed as follows:

$$\frac{A_c}{\sigma_c} = \frac{\sum_{i=1}^{M} k_i A_i}{\sqrt{\sum_{i=1}^{M} (k_i \sigma_{ei})^2}} \qquad [6]$$

We showed that optimal combination is realized when the weighting factors (**Appendix A**) are:

$$k_i = A_i / \sigma_{ei}^2 \qquad [5b]$$

In addition to the theoretically derived weighting factor, we also proposed an empirically derived weighting factor, which is the SNR_e itself:

$$k_i = A_i / \sigma_{ei} \qquad [5c]$$

We compared and evaluated the four weighting factors using phantom and in vivo data. We use σ to represent σ_e in the rest of the paper.

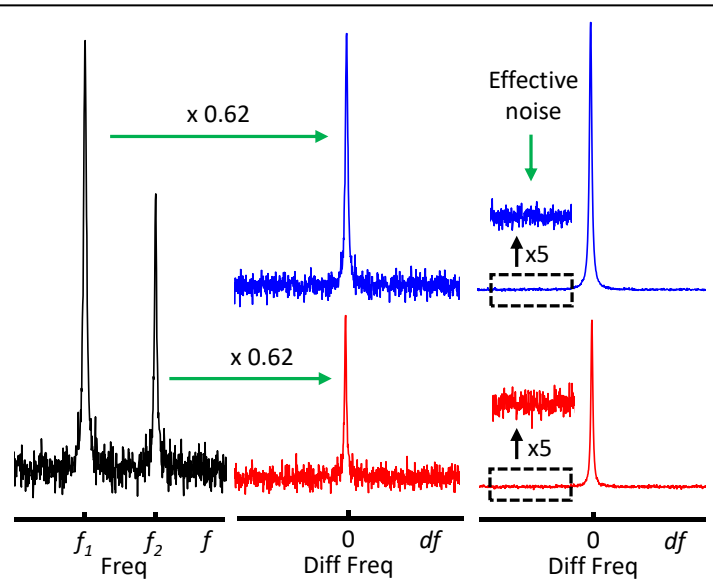

Figure 2. A simulated MRS spectrum (black, left-hand panel) with two component peaks, each of which has its own peak height, linewidth, and resonance frequency (f_1 or f_2), and with Gaussian noise. To facilitate the combination of the peaks at different frequencies, the two peaks are depicted on difference-frequency axis (df) with respect to their resonance frequencies of f_1 and f_2, respectively (center panel). Their noise levels are rescaled to the equivalent noise according to their CRLBs for frequency in Hz (Eq. 2b) to account for the effect of noise on the frequency measurement (right). Now the equivalent noise represents the achievable precisions of the frequency measurement of the individual peaks. Note that the df can be converted to temperature, with its origin assigned to a T_0, eg, 37°C (see the text for details).

2.1.5 Errors of averaged temperatures

The differential of the averaged temperature is derived from Eq. 4 as the following:

$$dT_c = \sum_{i=1}^{M}(k_i dT_i + T_i dk_i) \quad [7]$$

where T_i and dT_i are the temperature and its variation, and k_i and dk_i are the weighting factor and its variation of the i-th reference, respectively. This equation indicates that the errors of the averaged temperature result from the errors of temperatures of individual references and the weighting factors. In practice, repeated measurements or Monte Carlo studies are commonly employed to assess the errors of T_c can be measured. The j-th measurement is given as,

$$dT_{c,j} = \sum_{i=1}^{M}(k_{i,j} dT_{i,j} + T_{i,j} dk_{i,j}) \quad [8]$$

The standard deviation of $\{ dT_{c,j}\}$ is a metric of the errors of the averaged temperature, which is the same as that measured directly from $\{T_{c,j}\}$. We note that when both $dT_{i,j}$ and $dk_{i,j}$ are small among the repeated measurements, $T_{i,j}$ and $k_{i,j}$ can be replaced by $\bar{T}_i$ and $\bar{k}_i$, respectively (Appendix B). Therefore, Eq. 8 allows separating the contributions of errors in $T_{i,j}$and $k_{i,j}$ to the total errors of T_c.

2.2 ^{1}H MRS data acquisition and processing

2.2.1 Phantom data

Data acquisition We acquired single voxel phantom ^{1}H MRS data on a 3T scanner (SIGNA™ Premier, GE Healthcare) using a 21-channel surface coil and PROBE-P, a commercial PRESS sequence.[12] The data acquisition parameters were as follows: TR/TE = 2000/120 ms, spectral width = 5000 Hz, FID points = 4096, number of excitations for the unsuppressed water = 16, number of saved water FIDs = 2; number of excitations for water suppressed data = 240, numbers of saved, water-suppressed FIDs (each with an 8-step phase cycling) = 30, voxel size = 4x4x4 cm^3. Four MRS sessions were performed, each of which lasted 8 min.

Data preprocessing We combined the data from coil elements using unsuppressed water signal as a reference[13,14]. We removed the residual water signal using an SVD-based method[15-17] and performed spectral alignment among the 30 water suppressed FIDs by aligning the 2nd to the 30th FIDs to the 1st one, using a fitting procedure like the one by Near et al.[18]

Evaluation of the temperature averaging – by repeated measurements We used two methods to evaluate the performance of the temperature averaging. The first method used the repeated measurements, i.e., the 30 FIDs in each phantom MRS session. We fitted individual FIDs using a Lorentzian line shape model for the amplitudes $\{A_{i,j}\}$, frequencies $\{f_{i,j}\}$, decay rates $\{\alpha_{ij}\}$, and phases $\{\varphi_{ij}\}$, where i represents NAA, Cr, and Ch, respectively, and j = 1 to 30 represents individual FIDs. We also calculated the original noise levels from the FIDs and converted them to the equivalent noise leveles σ_{ij}, using the measured amplitudes, decay rates α_{ij}, according to Eq. 3. We converted the unit of frequencies to ppm and derived the temperatures from individual references using the calibration factors given by Zhu et al[19]. Initial results using these calibrations showed that the temperatures derived from Ch were remarkably different from those

derived from NAA and Cr and that this further induced large errors in the average temperatures. We therefore modified the calibration factors so that the temperatures measured from the three references, NAA, Cr, and Ch, were closest for the three sets of phantom data used. The temperature equations with modified interceptions are as follows:

$$\begin{aligned} T_{NAA} &= 313.9090 - 103.80 \cdot (\delta_w - \delta_{NAA}) \\ T_{Cr} &= 204.5278 - 101.70 \cdot (\delta_w - \delta_{Cr}) \\ T_{Ch} &= 195.0871 - 106.08 \cdot (\delta_w - \delta_{Ch}) \end{aligned} \tag{9}$$

where δ_w is the frequency (in ppm) of the unsuppressed water signal. We measured temperatures from individual references $T_{r,j}$ and calculated the averaged temperatures $T_{c\,j,k}$, where i = {NAA, Cr, Ch}, j = 1, 2, …, 30, and $k = \{A, A^2, A/\sigma_e, A/\sigma_e^2\}$. We calculated the standard deviations of the $T_{r,j}$ and $T_{c\,j,k}$. As an example, the latter is given as follows:

$$SD_{T_{c_k}} = \sqrt{\frac{\sum_{j=1}^{N}(T_{c\,j,k} - \overline{T_c}\,_k)^2}{N}} \tag{10}$$

where N = 30. The optimally averaged temperatures had a mean value closest to the temperature derived from the reference with highest equivalent SNR and smallest standard deviation.

Evaluation of the temperature averaging – Effects of noise levels and linewidths The second approach used Monte Carlo (MC) simulations to test the effects of linewidths and noise levels, respectively, on the precisions of the averaged temperatures using different weighting factors. To test the effects of linewidths, we multiplied the original 30 signals in each MRS session by Lorentzian line shape functions with linewidths of 4 and 10 Hz, respectively, and we also added a complex noise signal with Gaussian distribution to individual line broadened free induction decay (FID). The noise levels were derived from the original FIDs. We proceeded with the processing of these line broadened and noise-added signals for the average temperatures, $T_{c\,j,k,l}$, as described above, where l represents the linewidth. We calculated the standard deviations of the averaged temperatures for each $\{k, l\}$ and compared the effects of linewidths. The effects of noise levels were assessed following similar procedures as that for the linewidths. To test the effects of noise levels, we added noise signals with 4- and 10-times their original noise, respectively, to the 30 original signals and proceeded to measure and calculate the SDs of the individual and averaged temperatures.

2.2.2 In vivo data

Data acquisition We acquired single voxel MRS data from five human subjects on the same 3T scanner as for the phantom experiments. The protocols for human studies were approved by the IRB and informed consent was obtained from each subject before the MR scan. The single voxel MRS data were acquired from medial prefrontal cortex using a 21-channel surface coil and the PRESS[12] sequence with the following parameters: voxel size: 3.0 x 2.5 x 2.5 cm^3, TR/TE = 1500/120 ms, spectral width = 2000 Hz, FID datapoints = 1024, number of saved, non-water-suppressed FIDs = 2, number of saved, water-suppressed FIDs = 30; total MRS data acquisition time was approximately 6.5 minutes.

Data preprocessing of the in vivo data, which included combinations of data from element coils, residual water removal and spectral alignment, was carried out using the methods as described above. The procedures for spectral fitting and for the conversion of frequencies to temperature, also known as temperature calibration, were also performed using the methods described above. Specifically, the temperature calibration was performed for individual subjects so that the temperatures derived from the three references, NAA, Cr, and Ch, were approximately the same.

Evaluation of temperature averaging by Monte Carlo simulations We used all in vivo MRS datasets, one from each of the 5 subjects, as the basis signals in the MC simulations. We first determined the original noise level for each dataset by calculating the standard deviation of the data points in the second half of the FID, where the metabolite signals decayed out. We created N = 500 sets of complex noise signals with Gaussian distribution and with the same noise levels as in the corresponding original, basis signals. We added these individual noises to an original, basis FID to form a set of test signals and submitted them to the MC procedure. We calculated the SDs of the averaged temperatures, $T_{c\,k}$:

$$\sigma_{T_{c_k}} = \sqrt{\frac{\sum_{j=1}^{N}(T_{c\ j,k} - \overline{T_c}_{\ k})^2}{N}} \qquad [11]$$

where j = 1, 2, …, N is the index of the noisy signals, k represents the weighting factor (Eq. 5), and $\overline{T_c}_{\ k}$is the average of the averaged temperatures for weighting factor k.

2.3 Computer simulations

We performed computer simulation to aid error analysis of the averaging of temperature. We constructed a noise-free MRS spectrum consisting of 3 peaks with a Lorentzian line shape, mimicking NAA, Cr (CH_3), and Ch (see Table 5 for the parameters). We added to the signal a Gaussian noise, with a standard deviation σ_0. We generated N noise signals, each with the same data points p as the noise-free signal:

$$\{\sigma_{0,n}\}_{N\times p} = \sigma_0 \cdot randn(N,p) \qquad [12]$$

where $randn(N,p)$ is an N x p matrix with normally distributed random numbers of nominal standard deviation of 1. We added these noise signals to the noise-free MRS signal to form N noisy signals. We processed the noisy signals to calculate individual reference temperatures and the averaged temperatures with two different procedures: (1) by using the given spectral parameters and (2) by using parameters obtained from spectral fitting, respectively.

In the first procedure, we assigned the temperatures for the three references as

$$\begin{aligned} T_{i,n} &= T_{r,0} + \Delta T_{i,n} \\ i &= NAA, Cr, Ch \\ n &= 1, 2, \ldots, N \end{aligned} \qquad [13]$$

where$T_{i,0}$is the base temperature of the references, and $\Delta T_{r,n}$ is the temperature variations for the n-th test that were calculated according to the Cramer-Rao lower bound (Eq. 2 and Appendix B) from the given decay rates, amplitudes, and actual noise levels of the noisy signals:

$$\Delta T_{i,n} = \lambda C_{ppm,i,n} \qquad [14]$$

where $C_{ppm,I,n}$ is the CRLB for frequency in ppm (Eq. 2b) for the n-th signal, and λ is the frequency-temperature coefficient. In the computer simulation studies, we used λ = 100 ^{o}C/ppm[20] for all reference peaks.

We used the given values of the amplitudes, linewidths (see Table 5), and the actual noise level calculated as the standard deviation of $\sigma_{0,n}$ in Eq. 12 to calculate weighting factors (Eq. 5) for the averaging of temperatures. We repeated the temperature assignment and temperature averaging for the N signals and calculated their means and standard deviations.

In the second procedure, the amplitudes, frequencies and decay rates (linewidths) of the noisy signals were obtained from spectral fitting, as described previously for the phantom and in vivo data. We used these parameters to calculate the individual temperatures and the average temperatures; we then calculated their means and standard deviations.

To resolve and compare the contributions of errors of individual temperatures ΔT_i and weighting factors Δk_i to errors of the averaged temperature ΔT_c (Eq. 7), we intermingled (1) the theoretical individual temperatures used in the first procedure and fitted spectral parameters in the second procedure, and(2) the fitted individual temperatures in the second procedure and the given spectral parameters in the first procedure, respectively, and re-calculated the ΔT_c's.

3. Results

Phantom experiments – original data

Examples of the spectral fitting of the phantom data, including the original, the noise added, and the line broadened spectra, are given in Figure 3. The amplitudes, frequencies, and linewidths measured for the original 30 FIDs from experiment 1 were shown in Figure 4. The reference temperatures, T_{NAA}, T_{Cr}, and T_{Ch}, were calculated from the frequencies and the average temperatures of T_A and T_{A^2} were calculated from the amplitudes, and $T_{A/\sigma}$ and T_{A/σ^2} were calculated from both amplitudes and linewidths. The standard deviations of these temperatures in the repeated measurement serve as the metrics of the precisions of the temperature measurements including temperature averaging. In all four experiments (Table 1), the A/σ weighting performed best, and both A/σ and A/σ^2 weightings outperformed other weightings and individual temperature measurements. The A^2 weighting outperformed the A weighting in three experiments, but it was inferior to the best individual measurement of T_{NAA}.

Phantom experiments – with added noise

The results of the phantom experiments with added noise seemed trivial in that the observations about the ranking of the precisions of the temperature measurement with original data remained largely unchanged: The A/σ weighting is still the best, but the SD of T_c with A/σ^2 weighting is slightly lower than that of A/σ weighting for Exp 3 with 10 times added noise (Table 2). Specifically, noise levels of the signal did not meaningfully alter the relative relationship between the weightings A and A^2 and between A/σ and A/σ^2.

Phantom experiments – with line broadening

Line broadening had remarkable influences on the precisions of the averaged temperatures (Table 3). In all cases except Exp. 1 with 4 Hz line broadening, A/σ^2 outperformed A/σ and became the best weighting factor. But the A/σ weighting was still better than the other two weightings (A and A^2). The A^2 weighting remained superior to the A weighting and surpassed the best individual temperature measurement T_{NAA}.

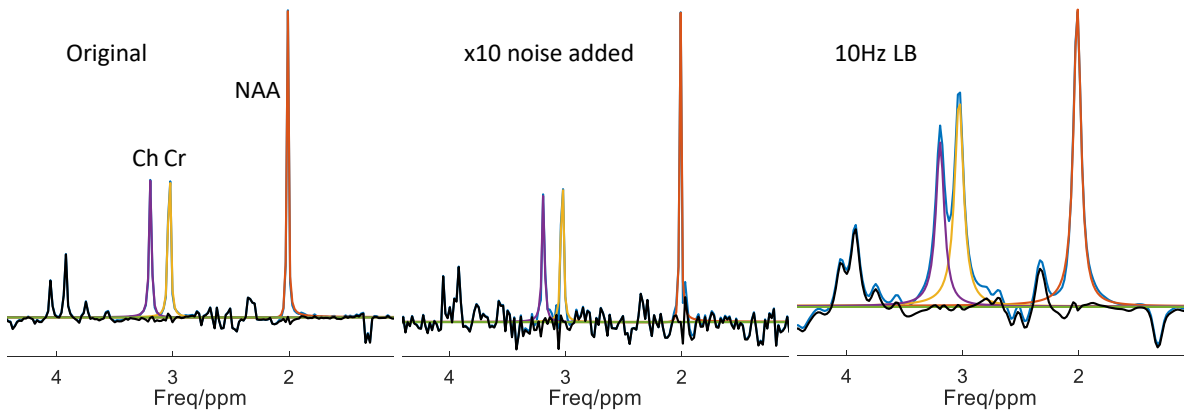

Figure 3. Examples of phantom spectral fitting for the original (left), x10 noise added (center), and 10 Hz line-broadened spectra (right). The spectra were fitted in the time domain using a Lorentzian line shape for NAA, Cr, and Ch. The blue and black lines are the signal to be fitted and the residue of the fitting, respectively. The other three colored lines are fitting spectra of NAA, Cr, and Ch.

Table 1. Standard deviations (SDs) of measurements of temperatures of individual references and their combinations with different weighting factors. Data were calculated from 30 FIDs in each experiment.

Exp.	Items	T_{NAA}	T_{Cr}	T_{Ch}	$T_{C,A}$	$T_{C,A}^2$	$T_{C,A/\sigma}$	$T_{C,A/\sigma 2}$
1	SD (x10^{-3})	0.793	4.119	6.506	2.178	1.742	0.581	0.635
	R2B	1.3649	7.0895	11.1979	3.7487	2.9983	1	1.0929
2	SD (x10^{-3})	2.138	4.888	9.154	1.973	2.485	1.803	1.99
	R2B	1.1858	2.7110	5.0771	1.0943	1.3783	1	1.1037
3	SD (x10^{-3})	1.328	5.046	6.588	2.189	1.809	1.027	1.150
	R2B	1.2931	4.9133	6.4148	2.1315	1.7614	1	1.1198
4	SD (x10^{-3})	1.4284	5.6820	7.7321	2.7732	2.2757	1.1484	1.1974
	R2B	1.2438	4.9479	6.7332	2.4149	1.9817	1	1.0427

Table 2. Standard deviations (SDs x 10^{-3}) of temperatures of the individual references and their weighted combinations with differing weighting factors. Data were from 4 experiments, each with 30 FIDs. Noises with 4x and 10x noise levels of the original FIDs were added, respectively.

Noise (x)	**Exp.**	T_{NAA}	T_{Cr}	T_{Ch}	$T_{C,A}$	$T_{C,A}^2$	$T_{C,A/\sigma}$	$T_{C,A/\sigma 2}$
4	**1**	7.36	19.46	32.84	11.09	9.54	6.81	7.04
	2	6.00	11.17	20.41	7.20	6.30	5.30	5.72
	3	9.22	25.48	41.90	12.83	10.92	8.10	8.60
	4	8.12	23.68	36.57	12.24	10.57	7.72	7.72
10	**1**	22.69	56.36	86.61	31.04	27.36	20.91	21.79
	2	20.11	55.38	88.44	29.55	25.63	18.40	19.10
	3	21.13	57.99	107.95	36.36	30.84	19.81	19.51
	4	22.35	60.44	93.85	30.36	26.40	20.51	21.16

Table 3. Standard deviations (SDs x 10^{-3}) of temperatures of the individual references and their weighted combinations with differing weighting factors. Data were from 4 experiments, each with 30 FIDs. The original FIDs were line broadened by 4 Hz and 10 Hz, respectively, for this test. To keep the reasonable noise levels for the data, Gaussian noises with standard deviations equal to the noise levels of the corresponding original FIDs were added to the FIDs after line broadening.

LB (Hz)	**Exp.**	T_{NAA}	T_{Cr}	T_{Ch}	$T_{C,A}$	$T_{C,A}^2$	$T_{C,A/\sigma}$	$T_{C,A/\sigma 2}$
4	**1**	8.41	14.55	20.69	8.26	7.69	7.15	7.24
	2	8.29	16.30	26.10	8.70	7.58	6.74	6.68
	3	17.80	32.63	39.66	16.60	15.46	14.62	14.42
	4	18.007	33.49	44.35	16.92	15.43	14.33	14.06
10	**1**	41.81	67.48	94.09	35.39	33.09	31.78	31.72
	2	46.91	94.47	124.75	47.50	44.35	42.28	40.18
	3	55.67	96.01	134.86	52.90	48.95	47.30	45.46
	4	59.20	88.44	131.71	52.56	48.87	47.12	45.88

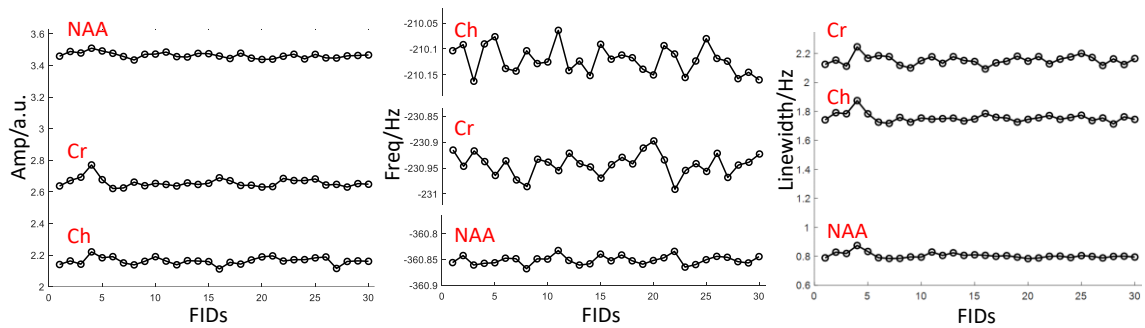

Figure 4. Examples of amplitudes (left), frequencies (center), and linewidths (right) of the three references measured from the 30 FIDs in an MRS session. The time interval between adjacent FIDs is 12 sec and the total time of lapse for the whole session is 6 min. Note that no frequency drift is seen over the period.

In vivo experiments

The spectral fitting of the in vivo data focused on the major signals of NAA, Cr, and Ch. Therefore, the J-coupled spectral peaks remained in the residue and they did not affect the quality of the spectral fitting (Figure 5). Of the data from the five subjects, the A/σ weighting outperformed other weighting factors in terms of the lowest SDs. The A and A/σ^2 weightings both had two 2nd-place rankings, but the former had three fourth-place rankings. The A^2 weighting had one 2nd-place and four 3rd-place rankings (Table 4). The SDs of all averaged temperatures were smaller than that of the best individual reference, which was NAA in this case, showing that all weighted averaging of temperatures improved the accuracy of temperature measurement. Overall, the proposed A/σ and A/σ^2 weightings was superior to the other previously proposed weightings.

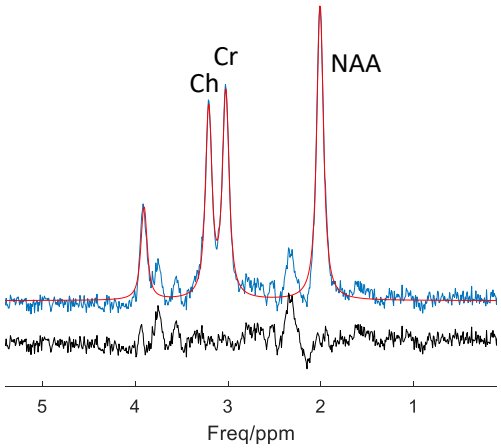

Figure 5. An example of spectral fitting of in vivo spectrum. The blue line is the spectrum to be fitted, the red line is the fitting spectrum, and the black line is the residue. Please note that Gaussian noise with the same level of noise in the original signal was added to the spectrum to facilitate the Monte Carlo study.

Table 4. Comparison of the accuracies of temperatures derived from individual references (NAA, Cr, and Ch) and different temperature combination algorithms (with weighting factors of A, A^2, A/σ, and A/σ^2). The original single voxel ^{1}H MRS were from five human subjects. Each MRS data set was combined with 1000 individual noise signals whose distribution is Gaussian and standard deviations are the same as those of the original signals. The above-mentioned individual temperatures from three references and averaged temperatures from four weighting factors were calculated and their standard deviations (SD) were presented here as a matric of the accuracy of the temperature measurement. R2B is "Ratio to the Best", representing the SD of a temperature to that of the best temperature measurement. The Rank is based on SDs.

Subj.	Items	T_{NAA}	T_{Cr}	T_{Ch}	$T_{C,A}$	T_{C,A^2}	$T_{C,A/\sigma}$	$T_{C,A/\sigma2}$
1	SD (x10^{-2})	8.2699	12.1774	8.7685	4.5802	4.5754	4.3746	4.5044
	R2B	1.8905	2.7837	2.0044	1.0470	1.0459	1	1.0297
	Rank	5	7	6	4	3	1	2
2	SD (x10^{-2})	8.2463	12.2642	9.2228	4.4705	4.4629	4.2816	4.4725
	R2B	1.9260	2.8644	2.1540	1.0441	1.0424	1	1.0440
	Rank	5	7	6	4	2	1	3
3	SD (x 10^{-2})	5.4464	8.6015	6.6135	3.3377	3.3332	3.2381	3.2729
	R2B	1.6820	2.6564	2.0424	1.0308	1.0294	1	1.0108
	Rank	5	7	6	4	3	1	2
4	SD (x 10^{-2})	10.1903	15.5803	13.2925	5.8028	5.9128	5.7561	6.0852
	R2B	1.7703	2.7067	2.3093	1.0081	1.0272	1	1.0572
	Rank	5	7	6	2	3	1	4
5	SD (x 10^{-2})	5.3373	8.8891	6.3851	3.9434	4.0351	3.8502	4.0370
	R2B	1.3862	2.3087	1.6584	1.0242	1.0480	1	1.0485
	Rank	5	7	6	2	3	1	4

Computer simulations

The parameters of the computer simulation are shown in Table 5. The SD of the Gaussian noise added to the signal was 0.1 (a.u.). The actual temperatures, amplitudes, and linewidths obtained from spectral fitting, together with the theoretical, given values, are shown in Figure 6. When the given signal parameters, actual noise levels (Eq. 12) and theoretically calculated temperature variations (Eq. 13) were used to calculate the weighting factors for the averaging of individual temperatures, the A/σ^2 weighting excelled other weightings by 1.1%, 4.8%, and 8.7%, respectively (Table 6). However, when fitting parameters for the noisy signals were used in the calculation of weighting factors and the measured temperatures were used in the averaging of the temperatures, the A/σ weighting performed best, surpassing A/σ^2, A^2, and A by 0.3%, 3.3%, and 7.3%, respectively.

In the intermingled experiments, the combination of theoretically calculated individual temperatures and spectral fitting obtained parameters provided almost the same average temperatures as the those in Procedure 1, indicating that the errors in the weighting factors, i.e., dk_i term in Eq. 8 and Appendix C8 played a minor role in the errors in the average temperature (Table 7); The combination of the individual temperatures from spectral fitting and the given spectral parameters resulted in almost the same average temperatures as those in Procedure 2, suggesting that the errors in the individual temperatures, i.e., dT_i term in Eq. 8 and Appendix C8, contributed a major part in errors of the average temperatures.

Table 5. Parameters used in computer simulation. The SD (σ_0) of the Gaussian noises added to the signal was 0.1 (a.u.). The number of tests is 1000.

*) Temperatures of the references *Temp* were set to 37 °C, and their error levels *ΔT were* given in the procedure using given spectral parameters and σ_0, but they and other parameters (Amp. etc) were obtained from the spectral fitting as shown in Figure 6.

	NAA	Cr	Ch
Amp. (a.u.)	7.6	7.0	6.0
Freq. (Hz)	-340	-212	-190
Linewidth (Hz)	7	9	8
Phase	0	0	0
Temp (°C)*	37	37	37
ΔT (°C, 10^{-2})*	1.0754	1.9860	1.4266

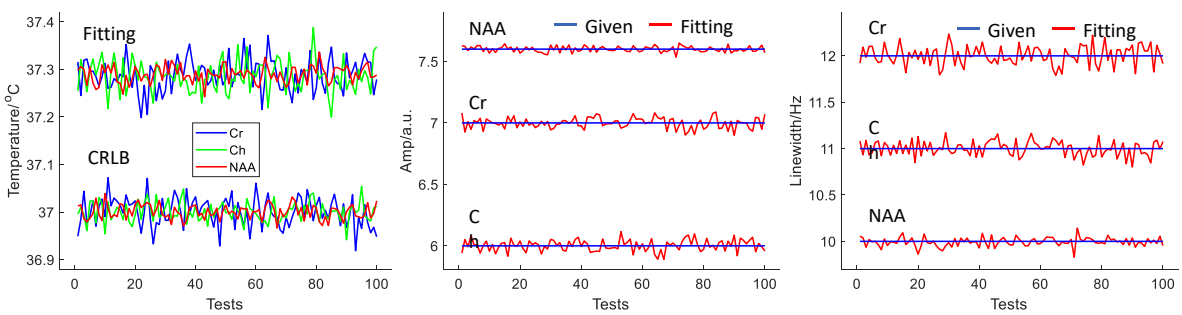

Figure 6. Temperatures, amplitudes, and linewidths used in the computer simulations. In the left panel, the lower traces were the theoretical temperatures calculated using Eq. 13, and the upper traces were temperatures obtained from spectral fitting. In the middle and right panels, the blue lines were the given values and red lines were obtained from spectral fitting. The number of tests was 1000 but only the first 100 tests were shown

Table 6. Standard deviations (SD) of individual and averaged temperatures in a computer simulation. The results were obtained from the two procedures: (Procedure 1) using given spectral parameters (amplitudes and linewidths) and theoretically calculated individual temperatures and (Procedure 2) using individual temperatures and spectral parameters from spectral fitting, respectively. R2B is "Ratio to the Best".

Proc.	Item	T_{NAA}	T_{Cr}	T_{Ch}	$T_{C,A}$	$T_{C,A}^2$	$T_{C,A/\sigma}$	$T_{C,A/\sigma 2}$
Given	**SD (x10^{-3})**	10.8416	20.4318	14.0770	8.6135	8.3058	8.0074	7.9229
	R2B	1.3684	2.5788	1.7767	1.0872	1.0483	1.0107	1
Fitting	**SD (x10^{-3})**	10.9239	23.6023	16.8657	8.3130	8.0007	7.7454	7.7703
	R2B	1.4104	3.0473	2.1775	1.0733	1.0330	1	1.0032

Table 7. Standard deviations (SD) of averaged temperatures in a computer simulation. Results were obtained using the data of (1) theoretically calculated individual temperatures and fitting-obtained spectral parameters (amplitudes and linewidths) and (2) fitting-obtained individual temperatures and given spectral parameters, respectively. R2B is "Ratio to the Best".

Data	Item	$T_{C,A}$	$T_{C,A}^2$	$T_{C,A/\sigma}$	$T_{C,A/\sigma 2}$
Theor. Temp + Fitted parameters	**SD (x10^{-3})**	8.6128	8.3047	8.0079	7.9242
	R2B	1.0869	1.0480	1.0106	1
Fitted Temp. + Given parameters	**SD (x10^{-3})**	8.3139	8.0024	7.7450	7.7686
	R2B	1.0733	1.0330	1	1.0032

Discussion

We have proposed concepts of equivalent noise and equivalent signal to noise ratio in terms of the measurement of frequency according to the CRLBs and a concept of relative frequency. Based on these concepts, we derived a theoretical weighting factor, A/σ^2, where σ is the equivalent noise, for the combination of relative frequencies of multiple references that may improve precision of temperature measurement. We also proposed an empirical weighting factor A/σ, which is the equivalent SNR. We carried out phantom and in vivo experiments to evaluate the performances of the two weighting factors and compared them with two previously proposed weighting factors. We also performed computer simulation studies to aid error analysis of different weighting factors. The results of computer simulation agreed reasonably well with those predicted by the theories, and the results of phantom and in vivo experiments showed superiority of the two proposed weighting factors over the previously proposed ones in terms of the precision of temperature measurement.

The concept of the equivalent noise, which involves not only the conventional spectral noise but also the decay rate (i.e., linewidth in the frequency domain), plays a fundamental role in the development of the two proposed weighting factors. First, it makes the two weighing factors peak specific. The conventional spectral noise is global, meaning that it is the same for all components (peaks) in the signal. Therefore, replacing the equivalent noise with the conventional spectral noise in the currently proposed weighting factors is meaningless, as they will be cancelled out in the normalization; and it is the reason that only A and A^2 weightings were previously suggested without involving noise[10]. Second, the equivalent noise is directly related to the measurement precision of frequency – a larger equivalent noise means lower precision of frequency measurement. Therefore, it is intuitive to place the equivalent noise or its square to the denominator of a weighting factor: a peak with larger equivalent noise should have smaller weighting in the combination, and vice versa.

The concept of relative frequency also played an important role in the development of the two proposed weighting factors. This concept makes different peaks in a spectrum equivalent or at the same location in the relative frequency axis, thus making possible the combination of peaks with different frequencies in a spectrum. The physical basis of the relative frequency in this paper is that the frequencies of different reference peaks correspond to the same temperature in ^{1}H MRS-based thermometry. In this sense, these frequencies are equivalent, and this is reflected in the relative frequency axis.

A/σ^2 weighting outperformed other weightings only in the phantom experiments with line broadening, and the results of both other phantom experiments (original data and noise-added data) and in vivo experiments showed that A/σ, instead of the theoretically optimal A/σ^2, is the optimal weighting factor. We note that the derivation of the optimal weighting A/σ^2 (**Appendix A**) is based on the optimal conditions that both the amplitudes and the decay rates did not have errors. In this case, the weighting factors can be correctly calculated. In real world data, errors in the weighting factors $k_{i,j}$ stem from measurement errors of amplitudes and decay rates (Eqs. 3 and 6). These errors may accumulate and propagate to affect the equivalent noise used in the proposed weighting factor, making them deviate from the correct values. The increased measurement errors in the individual temperatures as compared to the theoretically calculated ones, shown in the computer simulations, supported this analysis: the error of the individual temperature (**Appendix B**) is directly associated with error of the equivalent noise (Eq. 3), whose error will in turn spoil the proposed weighting factors. A/σ^2 weighting performed best when the given, noiseless amplitudes and decay rates (linewidths) were used in calculating the weighting factors, and it may be inferior to A/σ weighting because of its larger errors in σ^2 than in σ ($\sim \alpha^{1.5}/A$). This also explains why A weighting is better than A^2 weighting in some cases because the Δk error in A^2 is double that in A.

The optimal averaging of temperatures is similar to the combination of MRS data acquired using multichannel coil arrays, where equal weighting[21], amplitude weighting[22, 23], signal to noise (S/N) weighting[14], and signal to squared noise (S/N^2) weighting[21] were proposed. While the S/N^2 weighting was theoretically derived as the optimal weighting factor, a study[21] showed that its performance may be inferior to the S/N weighting. In fact, the situation here is more complex than the combination of the multichannel coil MRS data. In the latter, only amplitudes of water signal and spectral noises are needed, and both can be measured more accurately than the amplitudes of the reference peaks and equivalent noise. In the present case, the spectral noises, linewidths, and much weaker signal amplitudes of metabolites than water are needed. Because the A/σ^2 involves the square of the equivalent noise σ, the errors caused by the errors in amplitudes, decay rate, and spectral noises would be larger than errors in A/σ. Therefore, the A/σ^2 weighting is prone to be inferior to the A/σ weighting, more than is the case for multichannel coil data combination.

As analyzed in the Methods section and the Appendix A, a simple averaging of frequencies/temperatures may not ensure improved precision, i.e., the precision of the average temperatures may not surpass the precision of the temperature from the best individual reference. In the experiments using original phantom data, the SDs of both A and A^2 weightings were inferior to that of NAA reference. This has been shown in a previous study[10], where the SDs of all averaged temperatures using A^2 weighting were smaller than those of single referenced T_{NAA} and T_{Cr} but larger than that of T_{Ch}, which was derived from the dominant peak of Ch. Recently[24] Maudsley et al also found no improvement using A^2 weighting compared with the best individual reference.

Conclusion

We proposed concepts of equivalent noise, equivalent SNR, and relative frequency in terms of frequency measurement and the combination of peaks of different frequencies. Based on these concepts, we derived a theoretically optimized weighting factor and proposed an empirical weighting factor for the averaging of temperatures measured from three references. Experiments using phantom and in vivo data showed that these two weightings outperformed previously

proposed weightings in improving the temperature measurement using the ^{1}H MRS-based thermometry.

Reference

1. Odeen H, Parker DL. Magnetic resonance thermometry and its biological applications - Physical principles and practical considerations. *Prog Nucl Magn Reson Spectrosc* 2019; **110:** 34-61.

2. Bainbridge A, Kendall GS, De Vita E, Hagmann C, Kapetanakis A, Cady EB *et al.* Regional neonatal brain absolute thermometry by 1H MRS. *NMR Biomed* 2013; **26**(4)**:** 416-423.

3. Babourina-Brooks B, Wilson M, Arvanitis TN, Peet AC, Davies NP. MRS water resonance frequency in childhood brain tumours: a novel potential biomarker of temperature and tumour environment. *NMR in biomedicine* 2014; **27**(10)**:** 1222-1229.

4. Marshall I, Karaszewski B, Wardlaw JM, Cvoro V, Wartolowska K, Armitage PA *et al.* Measurement of regional brain temperature using proton spectroscopic imaging: validation and application to acute ischemic stroke. *Magn Reson Imaging* 2006; **24**(6)**:** 699-706.

5. Zhu M, Sun Z, Ng CK. Image-guided thermal ablation with MR-based thermometry. *Quant Imaging Med Surg* 2017; **7**(3)**:** 356-368.

6. Posporelis S, Coughlin JM, Marsman A, Pradhan S, Tanaka T, Wang H *et al.* Decoupling of Brain Temperature and Glutamate in Recent Onset of Schizophrenia: A 7T Proton Magnetic Resonance Spectroscopy Study. *Biol Psychiatry Cogn Neurosci Neuroimaging* 2018; **3**(3)**:** 248-254.

7. Kohut SJ, Kaufman MJ. Magnetic resonance spectroscopy studies of substance use disorders: Current landscape and potential future directions. *Pharmacol Biochem Behav* 2021; **200:** 173090.

8. Xu D, Vigneron D. Magnetic resonance spectroscopy imaging of the newborn brain--a technical review. *Semin Perinatol* 2010; **34**(1)**:** 20-27.

9. Horska A, Barker PB. Imaging of brain tumors: MR spectroscopy and metabolic imaging. *Neuroimaging Clin N Am* 2010; **20**(3)**:** 293-310.

10. Cady EB, Penrice J, Robertson NJ. Improved reproducibility of MRS regional brain thermometry by 'amplitude-weighted combination'. *NMR Biomed* 2011; **24**(7)**:** 865-872.

11. Hartmann J, Gellermann J, Brandt T, Schmidt M, Pyatykh S, Hesser J *et al.* Optimization of Single Voxel MR Spectroscopy Sequence Parameters and Data Analysis Methods for Thermometry in Deep Hyperthermia Treatments. *Technol Cancer Res Treat* 2017; **16**(4)**:** 470-481.

12. Bottomley PA. Spatial localization in NMR spectroscopy in vivo. *Ann N Y Acad Sci* 1987; **508:** 333-348.

13. Lin Y-Y, Hodgkinson P, Ernst M, Pines A. A Novel Detection – Estimation Scheme for Noisy NMR Signals:Applications to Delayed Acquisition Data. *JOURNAL OF MAGNETIC RESONANCE* 1997; **128:** 30-41.

14. Dong Z, Peterson B. The rapid and automatic combination of proton MRSI data using multi-channel coils without water suppression. *Magn Reson Imaging* 2007; **25**(8)**:** 1148-1154.

15. Dong Z, Dreher W, Leibfritz D. Toward quantitative short-echo-time in vivo proton MR spectroscopy without water suppression. *Magn Reson Med* 2006; **55**(6)**:** 1441-1446.

16. Kantrowitz JT, Dong Z, Milak MS, Rashid R, Kegeles LS, Javitt DC *et al.* Ventromedial prefrontal cortex/anterior cingulate cortex Glx, glutamate, and GABA levels in medication-free major depressive disorder. *Transl Psychiatry* 2021; **11**(1)**:** 419.

17. Dong Z, Grunebaum MF, Lan MJ, Wagner V, Choo TH, Milak MS *et al.* Relationship of Brain Glutamate Response to D-Cycloserine and Lurasidone to Antidepressant Response in Bipolar Depression: A Pilot Study. *Front Psychiatry* 2021; **12:** 653026.

18. 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. *Magn Reson Med* 2015; **73**(1)**:** 44-50.

19. Zhu M, Bashir A, Ackerman JJ, Yablonskiy DA. Improved calibration technique for in vivo proton MRS thermometry for brain temperature measurement. *Magn Reson Med* 2008; **60**(3)**:** 536-541.

20. Wang P. Evaluation of MR thermometry with proton resonance frequency method at 7T. *Quant Imaging Med Surg* 2017; **7**(2)**:** 259-266.

21. Hall EL, Stephenson MC, Price D, Morris PG. Methodology for improved detection of low concentration metabolites in MRS: optimised combination of signals from multi-element coil arrays. *Neuroimage* 2014; **86:** 35-42.

22. Natt O, Bezkorovaynyy V, Michaelis T, Frahm J. Use of phased array coils for a determination of absolute metabolite concentrations. *Magn Reson Med* 2005; **53**(1)**:** 3-8.

23. Wijtenburg SA, Knight-Scott J. Reconstructing very short TE phase rotation spectral data collected with multichannel phased-array coils at 3 T. *Magn Reson Imaging* 2011; **29**(7)**:** 937-942.

24. Maudsley AA, Goryawala MZ, Sheriff S. Effects of tissue susceptibility on brain temperature mapping. *Neuroimage* 2017; **146:** 1093-1101.

Appendix A: Show that $\{k_i = A_i / \sigma_i^2\}$ are the optimized weighting factors for the weighted averaging of multiple frequency measurements.

We start from an intuitive example of two peaks with (A; σ) = (40; 10) and (10; 1), respectively. First, we let $\{k_i = A_i\}$, which means the weighting factors are proportional to their corresponding amplitudes. The combined peak is (34; 8.0), whose *SNR* is $R = 4.25$ – larger than the first but small than the second peak. The reason for the failure of this weighting is that it does not take the noise into account.

Now we let $\{k_i = A_i / \sigma_i\}$, meaning that the weighting factors are proportional to their $\{R\}$. Using the above example, we obtained the combined peak of (18.57; 2.95), whose R is 6.31 but is still smaller than the second one. The reason for the failure is that the SNR is not normalized but proportional to the noise. To overcome the problem, we use noise level to normalize SNR and let $\{k_i = A_i / \sigma_i^2\}$. The combined peak is given by (11.15; 1.04), whose R is 10.77. To derive an optimized weighting, we assume $k_i = A_i / \sigma_i^x$ and determine an optimal x.

Substituting $k_i = A_i / \sigma_i^x$ into Equation (4), we have:

$$\frac{A_c}{\sigma_c} = \frac{\sum_{i=1}^{n} A_i^2 \sigma_i^{-x}}{\sqrt{\sum_{i=1}^{n} A_i^2 \sigma_i^{2-2x}}} \qquad [A1]$$

where A_i and σ_i are amplitude and noise level of the i-th peak, respectively.

Taking the derivative of the above equation with respect to x, we obtain:

$$\frac{d}{dx}\left(\frac{A_c}{\sigma_c}\right) = \frac{1}{\sqrt{\sum_{i=1}^{n} A_i^2 \sigma_i^{2-2x}}}\left[\frac{\sum_{i=1}^{n} A_i^2 \sigma_i^{-x} \sum_{i=1}^{n} A_i^2 \sigma_i^{2-2x} \ln\sigma_i}{\sum_{i=1}^{n} A_i^2 \sigma_i^{2-2x}} - \sum_{i=1}^{n} A_i^2 \sigma_i^{-x} \ln\sigma_i\right]. \qquad [A2]$$

Solving

$$\frac{d}{dx}\left(\frac{A_c}{\sigma_c}\right) = 0 \qquad [A3]$$

gives $x = 2$ and the optimized frequency measurement is

$$\frac{A_c}{\sigma_c} = \frac{\sum_{i=1}^{n} A_i^2 \sigma_i^{-2}}{\sqrt{\sum_{i=1}^{n} A_i^2 \sigma_i^{-2}}}. \qquad [A4]$$

For the special case where all measurements have the same A_i/σ_i, the above equation reduces to its well-known form for signal accumulation:

$$\frac{A_c}{\sigma_c} = \frac{A_i}{\sigma_i}\sqrt{n}. \qquad [A5]$$

Appendix B: Temperature errors due to frequency errors of the references

The temperature measured from the frequencies of water and the reference is given as follows:

$$T_r = \lambda(\delta_r - \delta_w) + T_{int} \qquad \text{[B1]}$$

where T_r is the temperature measured from a reference, eg, the singlet of NAA at 2.01 ppm, λ is the frequency-to-temperature coefficient, δ_r and δ_w are the frequencies of the reference and water, respectively, and T_{int} is the intercept. Both λ and T_{int} are constants determined by the calibration experiment.

The error of T_r is:

$$\Delta T_r = \lambda(\Delta\delta_r - \Delta\delta_w) \qquad \text{[B2]}$$

According to the CRLB (Eq. 2), the measurement error of water frequency $\Delta\delta_w$is more than 3 order of magnitude smaller than of the reference $\Delta\delta_r$ and, therefore, can be ignored. This results in:

$$\begin{aligned} \Delta T_r &= \lambda\Delta\delta_r \\ &= \lambda C_{ppm,r} \end{aligned} \qquad \text{[B3]}$$

where $\Delta\delta_r$ is substituted by the Cramer-Rao low bound for frequency measurement, $C_{ppm,r}$.

Appendix C: Errors of the averaged temperature

For the reader's convenience, Eqs. 4 and 7 are copied here:

$$T_c = \sum_{i=1}^{M} k_i T_i \qquad \text{[C1]}$$

The differential of the averaged temperature is derived from the above equation as follows:

$$dT_c = \sum_{i=1}^{M} (k_i dT_i + T_i dk_i) \qquad \text{[C1]}$$

where T_i and dT_i are the temperature and its variation, and k_i and dk_i are the weighting factor and its variation of the i-th reference, respectively. In repeated measurements or in the Monte Carlo studies, $dT_{i,j}$ and $dk_{i,j}$ of the j-th measurement can be defined as,

$$\begin{aligned} dT_{i,j} &= T_{i,j} - \bar{T}_i \\ dk_{i,j} &= k_{i,j} - \bar{k}_i \\ j &= 1, 2, \ldots, N \end{aligned} \qquad \text{[C3]}$$

where N is the number of measurements or tests, and $\bar{T}_i$ and $\bar{k}_i$ are average temperature and average weighting factor, respectively, of the reference i in the repeated measurements. The averaged temperature from the j-th measurement is written as:

$$dT_{c,j} = \sum_{i=1}^{M} (k_{i,j} dT_{i,j} + T_{i,j} dk_{i,j}) \qquad \text{[C4]}$$

In this case, the standard deviation of $\{dT_{c,j}\}$can be used as a metric of the precision of the averaged temperature, which is equivalent to that calculated directly from the average temperatures of$\{T_{c,j}\}$.

Eq. C4 indicates that not only the differentials $dT_{i,j}$ and $dk_{i,j}$ but also the variations of $T_{i,j}$ and $k_{i,j}$ among repeated measurements contribute to the errors in the averaged T_c. We therefore term the $dT_{c,j}$ in Eq. 11 as total variation. To eliminate the effects of variations of $T_{i,j}$ and $k_{i,j}$, we define a new variation $dT_{c,j}^{m}$ as

$$dT_{c,j}^{m} = \sum_{i=1}^{M} (\bar{k}_i dT_{i,j} + \bar{T}_i dk_{i,j}) \qquad \text{[C5]}$$

The terms $\sum_{i=1}^{M} \bar{k}_i dT_{i,j}$ of Eq. C5 are variations stemming from weighted average of individual temperature differentials ($dT_{i,j}$), which arise from errors of frequency measurement (Eq. 2); The average weighting factors $\{\bar{k}_i\}$ do not contribute to these terms. The $\sum_{i=1}^{M} \bar{T}_i dk_{i,j}$ terms are the variations proportional to differentials of weighting factors, which arise from the errors in the measurement of the amplitudes and decay rates of the references (Eqs. 2 and 5); the average temperatures $\{\bar{T}_i\}$ do not contribute to these terms. Therefore, the partial variations $dT_{c,j}^{p}$ disentangle the contributions of errors of temperature measurements and the errors in the weighting factors.

From Eqs. C3 – C5 we obtain:

$$dT_{c,j} = dT_{c,j}^{m} + 2\sum_{i=1}^{M} dk_{i,j} dT_{i,j} \qquad \text{[C6]}$$

The above equation shows that the differences between $dT_{c,j}$ and $dT_{c,j}^{m}$ are second-order small quantities. In the cases when both $dT_{i,j}$ and $dk_{i,j}$ are small the second term can be ignored and Eq. C4 can be written as:

$$dT_{c,j} = \sum_{i=1}^{M}(\bar{k}_i dT_{i,j} + \bar{T}_i dk_{i,j}) \quad \text{[C7]}$$

When $\bar{T}_1 = \bar{T}_2 = \ldots = \bar{T}_M = T$, C7 becomes,

$$dT_{c,j} = \sum_{i=1}^{M} \bar{k}_i dT_{i,j} + T \sum_{i=1}^{M} dk_{i,j} \quad \text{[C8]}$$

The normalization condition of the weighting factor $\sum_1^M k_i = 1$ (Eq. 4b) ensures:

$$\sum_{i=1}^{M} dk_i = 0 \quad \text{[C9]}$$

We finally have:

$$dT_{c,j} = \sum_{i=1}^{M} \bar{k}_i dT_{i,j} \quad \text{[C10]}$$

C10 shows that the errors in the average temperatures are mainly caused by the errors in the individual temperatures.