# Clinical Trials of MRS Methods

Margarida Julià-Sapé [1, 2], Fernando Arias-Mendoza[3] and John R Griffiths[4]

[1] Centro de Investigación Biomédica en Red (CIBER). Cerdanyola del Vallès. 08193. Spain.

[2] Universitat Autònoma de Barcelona (UAB). Departament de Bioquímica i Biologia Molecular and Institut de Biotecnologia i Biomedicina. Cerdanyola del Vallès. 08193. Spain.

[3] Department of Radiology, University of Pennsylvania Health System, Philadelphia Pennsylvania 19104, USA

[4] Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE UK

## 1. ABSTRACT

In order to determine the applicability of noninvasive magnetic resonance spectroscopy (MRS) to the study of a diseased tissue or organ in the human body, it is necessary to determine if MRS is safe and effective. This is the primary purpose of a clinical trial. A clinical trial for MRS may also reveal which technical approach works best for the specific application and characteristics of the population being studied. In this chapter, we discuss the legal, ethical, and scientific requirements to be considered prior to the start of a clinical trial of an MRS protocol, as well as constraints that may arise during its execution. MRS-specific issues arising from a couple of successful clinical MRS trials for classifying brain tumors with $^1$H MRS (INTERPRET and eTUMOUR) and body tumors with $^{31}$P MRS (the Cooperative Group on MRS Application in Cancer, CoGMAC), serve as illustrative examples.

## 2. INTRODUCTION

The clinical value of MRS, like that of any other medical test, must be assessed in clinical trials (i.e. research studies that prospectively assign human participants into health-related interventions to evaluate health outcomes, conducted according to formal rules). This chapter will discuss the issues

that need to be addressed when planning and conducting a clinical trial of an MRS method. We will also briefly review some of the multi-center trials that have so far been performed and lessons that can be learned from them. We will use examples from two multi-center studies aimed at classifying different brain tumor types, INTERPRET (1) and eTUMOUR (2), to illustrate $^{1}$H MRS, and one from the Cooperative Group on MRS Application in Cancer (CoGMAC), which performed several trials on patients of non-Hodgkin's lymphomas (NHL), sarcomas of soft tissue and bone, breast carcinomas and head and neck carcinomas, to illustrate $^{31}$P MRS (3,4).

## 3. METHODS USED IN CLINICAL TRIALS OF MRS

### 3.1. Preliminary considerations

A clinical trial of an imaging method must be conducted according to formal rules. Trials that do not follow such rules will be unlikely to meet the criteria for inclusion in the Evidence Based Medicine meta-analyses that are nowadays used to decide whether a new technique should be adopted or an old technique should be dropped. For instance, a study intended to demonstrate diagnostic accuracy of a novel method (the most common potential use for MRS) should follow the 25 criteria set out in the STARD initiative (STAndards for the Reporting of Diagnostic accuracy studies) (5).

Historically, very few published clinical MRS studies were designed to adhere to criteria such as STARD, or even to be formal clinical trials. Furthermore, most were performed at a single center on a small number of patients and thus were therefore lacking in statistical power. When formal meta-analyses of the clinical MRS literature were performed (reviewed by (6)) few of the many thousands of published clinical MRS papers were qualified for analysis, and those that met the Evidence Based Medicine criteria had often been performed on small groups of patients, or with earlier generation instruments. Consequently, MRS appeared to perform badly in the meta-analysis, and that poor performance eventually resulted in the denial of reimbursement for MRS (CPT 76390) across hundreds of healthcare

providers in the USA (see (6)). Table 1 shows a set of criteria devised by (6) that should be

followed if a trial is to be included in future EBM meta-analyses.

**Table 1 near here**

## 3.2. Logistical challenges

In a clinical trial of a medical device such as MRS, the patients studied should have specific

health conditions under clearly defined treatments. This requirement can be difficult to

satisfy. First, the definition of the health condition may change, or it could become subdivided

into two or more conditions. Similarly, a treatment considered the standard for a disease could

be changed during the trial for another that has proven to be more effective. A clinical trial of

a device cannot control such changes in medical practice, which may be adopted at different

times (or not at all) by the clinicians in the various collaborating centers.

Furthermore, in any multi-institutional clinical trial the logistical challenges related to time,

costs, patient tolerance, and involvement of personnel must be taken into account in the

project-planning phase. In addition, international cooperation for a trial adds another level of

complexity due to differences in the standards for diagnosis and treatment in each country.

The numbers of patients and MRI/MRS acquisitions must be clearly estimated and the time

frames for accrual as well as the planning for subsequent follow-up should be kept as short as

possible. Further, the inclusion of medical centers should be based not only in the hospital's

ability to perform the protocol but also in its ability to recruit patients, taking into account the

incidence rate of this particular disease and the population covered by the hospital involved.

Similarly, magnetic resonance (MR) technology including MRS has evolved in large strides

in the past two decades. For many years clinical trials were performed using the standard

clinical magnetic field strength of 1.5 Tesla, but in the past 10 years 3.0 Tesla clinical MR

scanners have become widely available and for the past five years whole-body MR scanners

at 7.0 Tesla have been tested in major clinical centers. Furthermore, other types of hardware

(i.e., single *vs.* phased-array coils) and software (i.e., single volume *vs.* multivolume

acquisitions; or the use of echo-planar algorithms to acquire MRS) continue to advance. An

informed decision should be made every time such technological advances become available in the middle of a clinical trial of MRS. One needs to evaluate carefully the benefits offered by the new technology against the requirements needed to test it, to adapt the trial to use it properly, and to properly combine the new data with the data acquired with the previously available technology. As with any medical advances, however, the managers of the trial may have little or no control over the adoption of technical improvements at the different collaborating centers.

In addition, plans must also be made for collection of additional data, such as clinical information, genomic data, images or reports from other imaging modalities or biopsies. If biological specimens (biopsies, fluids, blood) have to be collected, it is inevitable that collection will fail in some instances. The more data types and controls we want to collect, the less probable is that we will be able to collect matched datasets for all the types of information. For example, the eTUMOUR project collected data from 1317 patients, of whom 656 fulfilled the minimum requirements in terms of data availability, 470 fulfilled the quality requirements, and 289 fulfilled the clinical validation requirements. Only 222 cases fulfilled all three requirements at the same time (2).

The number and type of quality control (QC) procedures should be estimated as precisely as possible in any multicenter trial and particularly in an international one. Are these QC procedures already in place at all institutions or will they need to be developed and validated? Who will do that development and how long will it take? It is also important that all the crucial tasks should be performed by personnel who are under the control of the project's management. In a multidisciplinary patient evaluation, for instance, a study may have one consent form for the biopsy and another one for the MRS acquisition – both consents should be obtained by project personnel. If we are not planning a clinical trial managed by a contract research organization (CRO), enough trained human resources should be committed. Many important tasks in the study, such as development of a database, are not "scientific" but can still take years. Development and training time frames must be estimated

as realistically as possible.  In an ideal world all the participating staff would receive uniform training in the procedures to be used; in reality that is almost never possible because of time and financial constraints.

### 3.3. Legal regulations

Clinical trials are of course, subject to the laws of the country in which they are performed, and again multicenter trials performed in more than one country are made more complex because they have to obey several legal frameworks. For example, when the MRS files have to travel outside the originating clinical center for analysis, or to a database, they must be anonymized or de-identified by stripping any data allowing identification of the patient or volunteer. Anonymization was traditionally a bottleneck in clinical trials of MRS until the advent of the DICOM MRS formats, although even today some MRS vendors (e.g. General Electric) have not yet moved to the DICOM standard.

One of the most important legal steps is when the clinician or academic researcher in charge (who in some countries will be the medical doctor responsible for the patient, not the academic researcher) must seek approval from the local Ethics Committee or Institutional Review Board (IRB). That obtained, the study can then start recruitment of patients, each of whom must give informed consent. Patient consent involves two documents: the information sheet and the consent certificate, the latter must be signed by the patient or his/her legal representative. Both documents must be written in language understandable to the patient, and must be stored by the clinician in charge at the clinical center where the medical procedure is carried out. In the US an additional document relating to confidentiality is also signed at the time of consent.

When a multicenter, international trial is conducted, the items that will appear on the information sheet must be agreed among the participating centers and then adapted to the national languages and standard forms of the participating hospitals and countries. In addition and to maintain uniformity, research centers participating in international research

funded by federal institutions in the US must follow the rules and regulations of the US funding institution, even if the regulations in the country concerned are different.

3.4. **Consensus protocols**

Consensus protocols for acquisition, processing and validation will be required for the MRS methods as well as for the additional data that are to be acquired by the study. As a general rule, it is better to use "standard of care" rather than "state of the art" acquisition protocols, unless, of course, the aim of the trial is to test a state of the art protocol. Standard of care protocols can be more easily adopted by (or may be already in use in) many possible contributing centers, and it may also be possible to obtain retrospective data for pilot studies, so as to increase the numbers of data from uncommon medical conditions. However, a lot of MRS techniques are not standard of care. In the US, the only MRS modality that has reached standard of care level is $^1$H MRS obtained using single voxel spectroscopy. This means that obtaining multivolume (MV) $^1$H MRS (usually using the chemical shift or the Hadamard algorithms) and all types of non-$^1$H MRS (usually referred as multinuclear MRS) like $^{31}$P or $^{13}$C MRS are considered non-standard of care and their use adds to the complexity of the trial and difficulty of finding institutions with the proper infrastructure to acquire them.

*3.4.1.Acquisition Protocols*

To minimize variability, most MRS trials have studied only one protocol (3,7) and/or, used a single scanner model (7). This restricts the number of potential participating centers and may not allow recruitment of enough cases with rare medical conditions. One solution, adopted by the INTERPRET project, is to define "compatible acquisition protocols" (Table 2), by performing pilot studies. For example, in (1), using a retrospective set of patients with brain tumors who had been studied using single voxel (SV) $^1$H MRS at 1.5T, it was found that spectra obtained by different sequences (STEAM and PRESS), different echo times (TE; 20, 30, 31, and 32 ms), two different repetitions times (TR; 1600 and 2000 ms), and two scanner

brands (GE and Philips) were still compatible for a pattern recognition (PR) study that aimed

to classify common brain tumor types.

**Table 2 near here**

In (3,4) the standardization and QC protocols set-up for a trial to use the MV (3D) acquisition

of $^{31}$P MRS at 1.5 Tesla to study patients with several forms of cancer (mainly lymphomas

and sarcomas) were described. Even though two scanner brands were used (GE, Siemens) in

an international multicenter study, reproducibility was optimized by using standard protocols

and a common radio-frequency coil design. Quality control studies were performed by

studying leg or thigh muscles of healthy human volunteers.

*3.4.2. Processing Protocols*

The spectra must be processed uniformly, using a single program and set of processing

parameters, as shown in Table 3 for SV $^{1}$H MRS, or as described in (3) for MV $^{31}$P MRS.
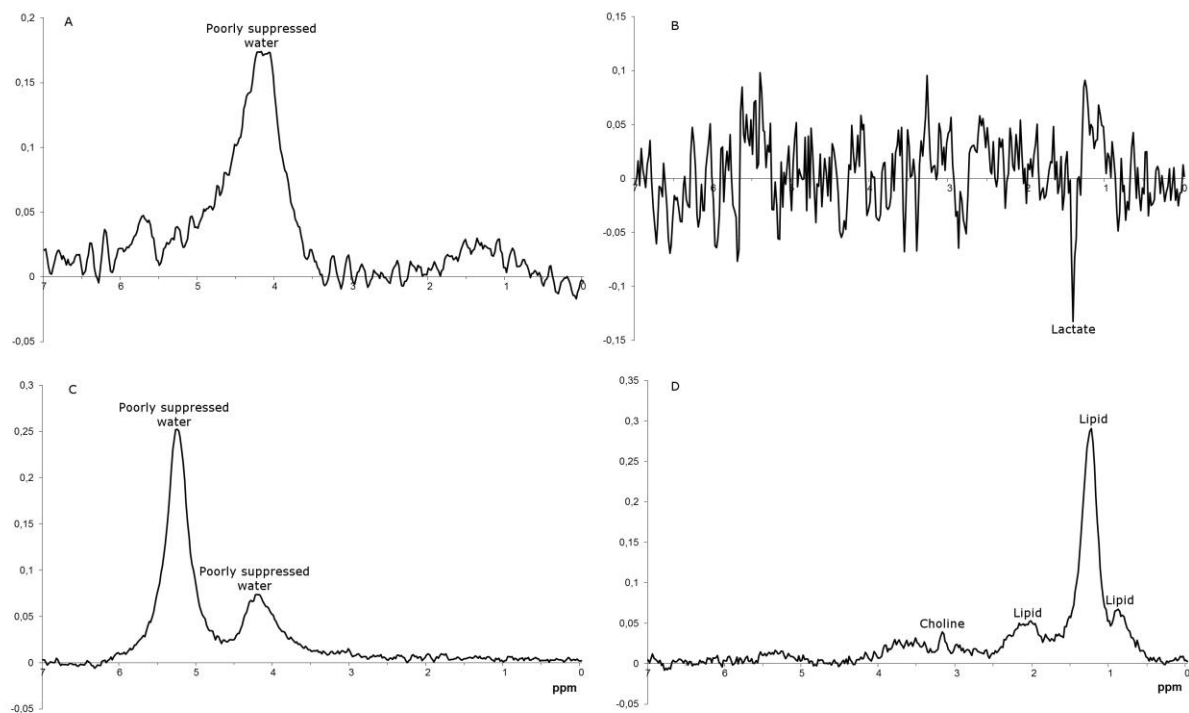
Table 3 and



Figure 1 show how automatic processing followed by data storage in a central database also

allows calculation of parameters such as signal-to-noise ratio (SNR) or water linewidth for

QC purposes, provides visualization software for the processed spectra and facilitates traceability of the MRS records. A disadvantage of automatic processing is the inevitable small percentage of lost data, as some spectra may be unusable without additional processing steps that are not included in the consensus protocol.

**Table 3 near here**

**Figure 1 near here**

Finally, the set of processing parameters must also include a common output format, such as the INTERPRET data manipulation software and canonical format (8), which has been adopted by some subsequent ^{1}H MRS trials (eTUMOUR and HealthAgents) (9). The particular feature of this canonical format is that each spectrum will have the same number of points along the ppm range, i.e. the same spectral resolution, which is particularly important when dealing with data obtained on several different scanner brands (8).

Although the spectra were processed manually using a proprietary program, the ^{31}P MRS trial in cancer (3,4) also used a standard set of processing parameters and a canonical format for the spectra.

An additional problem is normalization, particularly in pattern recognition (PR) studies where we are comparing differences in the relative proportions of multiple intensities, rather than quantifying the absolute areas of peaks. A simple approach to intensity normalization that has provided good results in several PR studies, is the Euclidean or $l_2$ norm (10), which is obtained by dividing the value of each individual point by the value of the norm. The norm is obtained by squaring the value of each point in the spectral vector, adding all them and taking the square root of this sum. The value of each point is then divided by the norm. This yields comparable spectra, without the need to take into account the water content of the tissue or the number of acquisitions, provided that all the spectra have the same number of points - hence the importance of a canonical format with standardized spectral resolution.

A different approach for normalization was introduced in the MV $^{31}$P MRS trial in cancer. In this case, for QC as well as for quantification and normalization purposes, a glass sphere of about 2 milliliters was placed in the center of the radiofrequency coil with a known solution of triphenylphosphite (TPP) that has a $^{31}$P signal with a T$_1$ value below 0.2 seconds. These standards were custom-made from the highest purity reagent available (Sigma Chemical Co., St. Louis, MO, USA). The $^{31}$P signal of TPP resonates approximately 3500 Hz away from the *in vivo* signals at 1.5 Tesla, allowing acquisition of its signal without affecting the *in vivo* results. The $90^0$ excitation pulse for TPP was calculated and using a TR of 1 second a fully-relaxed spectrum of 64 excitations was acquired for each human study. The area of the signal of TPP was integrated and its value used to normalized the *in vivo* signals (4).

### 3.4.3. Validation

Regular instrumental QC studies should be performed on phantoms and normal volunteers to ensure that each spectrum will have adequate SNR, good spectral resolution and absence of artifacts (3,4,11). In addition, every spectrum used for analysis should undergo an individual check, either by experts using standard criteria (4) or by PR algorithms using criteria previously set by experts (11,12). For example, the INTERPRET project used a PR approach on a subset of spectra to determine empirical thresholds for SNR and water linewidth; these thresholds were subsequently used for expert-based QC of individual spectra, both in INTERPRET and eTUMOUR (Table 4). If this semi-automated approach cannot be applied, which happened with the MV data in the INTERPRET project, the QC must be expert-based, using images of the data processed locally with the scanner software, and entered ad-hoc into the database by the data providers, along with relevant metadata such as field strength, sequence name, and the TE, TR, and voxels encoded for the X, Y and Z directions. The criteria for evaluation can be checked on Table 4. Manual data upload requires a significant input both from the data uploaders and the expert evaluators, and in the eTUMOUR project that resulted in a larger percentage of the manually uploaded MV spectra not completing their validation (65% of 613 spectra) as compared to SV (only 4.5% of 1653 spectra), in which

uploads were semi-automated. The eTUMOUR project also used HRMAS (high-resolution magic-angle spinning) spectra from brain tumor biopsies, and their QC was also entirely expert-based (Table 4), resulting in a wastage of 86% of the 1348 spectra because they did not have their QC completed. In contrast, microarray data upload was totally automated and only 0.3% of 665 cases were without a complete QC. These examples show how lack of automation could affect the validation of the massive amounts of data accrued in big multicenter trials.

**Table 4 near here**

The analysis of the ^{31}P MRS trial on human cancer has shown good reproducibility of the data acquired from different institutions (3) despite of being done without automation. In it, principal component analysis (PCA), was used instead of quantification to automate the analysis (13-15) processing and analysis of the ^{31}P MRS of non-Hodgkin lymphomas (NHL) (16). An advantage of this approach is that it does not require prior knowledge, but the approach is not suitable for datasets which contain spectra acquired under very different conditions (instruments, number of points in the FID, spectral bandwidth, etc.) (14).

The position of the volume or region of interest (ROI) must be verified and should accurately correspond with diagnostic or outcome variable information. In the case of single-voxel studies, accuracy needs to be ensured at the time of acquisition. Moreover, even though the volume selected might be correct, chemical shift displacement (see emrstm1481) may be a problem for SV as well as MV acquisitions that employ selective-excitation gradients in one or more dimensions (16). Thus voxel or ROI location must be carefully monitored: tumors are heterogeneous and, for instance, the spectrum will change if we acquire data from a nodular, cystic or edematous part of a brain tumor (Figure 3**Error! Reference source not found.**).

**Figure 3 near here**

MV acquisitions employing phase-encoding gradients are essentially devoid of this signal localization dependence on frequency in the phase-encoding dimensions, but more importantly allows the acquisition of signals from a large number of locations in a 2- or 3-dimensional volume, usually in a similar amount of time to single-volume techniques; furthermore, the location of the volume or region of interest can be readily adjusted after acquisition.(17)

Another source of data loss is disagreement among expert histopathologists and other clinicians, who provide the gold standard for diagnosis and characterization of the disease (18). This can cause a significant drop in the number of cases available (2,19): in INTERPRET about a third of the cases did not have an agreed histopathological diagnosis submitted. A similar number of cases were lost due to lack of clinical information in the $^{31}$P MRS trial of human cancers (4) .

In general terms, the more we can automate all aspects of validation the better, as it saves time and prevents both bottlenecks and expert bias. Centralized expert review should be avoided if possible, especially with respect to tasks that can be performed on-line. Nowadays the use of telepathology methods could eliminate the need for circulating histology slides by post to expert reviewers.

*3.4.4.Data storage.*

In MRS trials, data processing, data collection and analysis, as well as QC are usually centralized (3,7,20) and performed by experts at a coordinating institution. In contrast, in trials of pattern recognition MRS (2,8) (see section 2.6), we find the opposite: a centralized database that allows asynchronous and remote data entry, editing, download and monitoring. Although the development of an automated database is challenging, if the QC protocols and measurements are properly implemented and coupled to automated processing and display (e.g. MRS visualization tools), these processes can be performed remotely, saving expert time and avoiding bottlenecks. A centralized database can also store metadata, e.g. the instrument manufacturer, magnetic field, format, or SNR of each MR spectrum. This approach has been

successfully used (2,19) However, universal, automated, offline MRS data processing and display is still not possible, particularly for MV 2D and 3D MRS as well as their co-registration with MRI.

It is important to have all the data collected by the trial stored in the trial's database instead of discarding cases that do not fulfil inclusion criteria, since partially-compliant data can be used in secondary analyses or pilot studies. In a database, therefore, one does not exclude data, but rather established metrics for the number of "usable" cases and the percentage of cases not passing the QC thresholds for each data type. Such metrics allow the creation of subsets of data to answer different questions, ranging from data monitoring to analysis.

A further refinement is to have a distributed network of databases at each clinical center, each of which can be searched by the other centers. This approach was tested during the HealthAgents project  (21), and despite its attractiveness, disadvantages became apparent. Extensive security measures had to be implemented to prevent leaks of sensitive patient data; also the hosting and maintenance costs for the whole system were higher than for a centralized database.

## 2.5 Obstacles that inhibit clinical trials of MRS

There are two main obstacles that inhibit the development of clinical trials of MRS.  The first is the lack of any tradition for sharing data as a prerequisite for publishing MRS studies. As a consequence, there is a lack of consensus concerning data acquisition methods, processing protocols and output formats, particularly for MV data. There is also a lack of universal (either current or legacy formats), automated and standard methods for data processing, visualization and exchange, although programs such as 3D-*i*-CSI (22,23), LCmodel (24) and jMRUI (25) are able to process many formats, and ultimately SIVIC (26) can deal with some DICOM implementations.

The second obstacle is the rapid technological advance in MR mentioned in the "logistic challenges" section. Accrual of large multicenter datasets requires years (3,8), and for this, a

stable set of machines and sequences is needed: the benefit of using newer techniques for MRS acquisition need to be evaluated against the problems that will rise by modifying the method of observation in the middle of a trial. Of course, one could argue that a useful MRS method has to be able to function across a wide range of instruments, operating systems and pulse sequences, and the PR methods developed during the INTERPRET project, for instance, were able to "ignore" such technical differences and still successfully classify the tumor spectra. In most clinical trials, however, one will be attempting to minimize technical differences between the instruments and their software in order to test a hypothesis related to a medical condition.

## 2.6. Special issues in multicenter trials involving pattern recognition (PR)

Trials that incorporate PR methods in the analysis phase have particular features that determine their design and logistics. First, one needs a representative, non-biased sample of the population or disease of study, which is the main reason for scaling up to a multicenter trial. As in many biomedical domains, this need encounters two constraints: the "curse of dimensionality" and the "curse of data sparsity" (27). The number of features (points) in an MR spectrum will be in the thousands whereas the number of samples will typically be around 10-30 in the case of single-center studies on cancer. For instance, glioblastoma, the most common brain tumor, has an annual incidence-rate of 3.19 per 100,000 according to CBTRUS (28), oligodendroglioma is in the 0.26 per 100,000 range and lymphoma in the 0.44 per 100,000 range. The 304 cases in the multicenter INTERPRET database (19) include SV MRS spectra at short TE of 86 glioblastomas, 7 oligodendrogliomas, and 10 lymphomas, the latter two illustrating the curse of data sparsity.

The second requirement is similar to any other trial involving MRS. A series of controls must be performed to ensure that the acquired data are representative of the study population from both instrumental and clinical quality points of view. In particular, the datasets must not
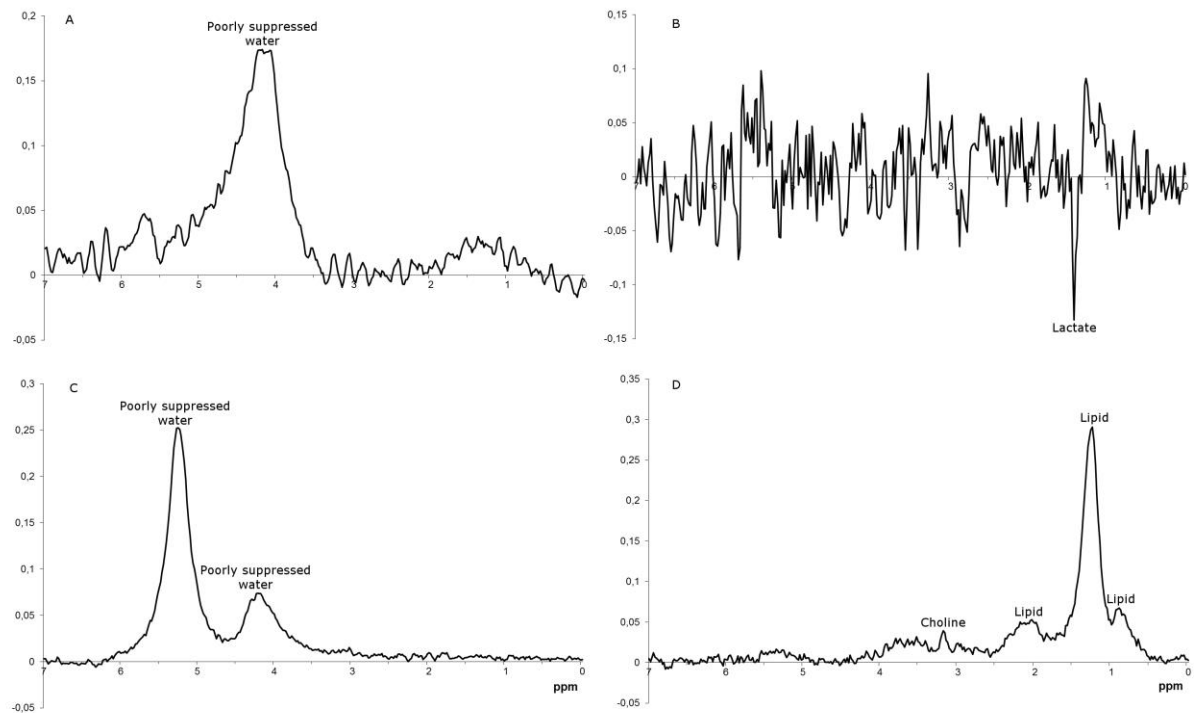
include any artefactual data (



Figure 1) or wrongly labelled cases (**Error! Reference source not found.**Figure 3.) that would add unnecessary noise to the natural variability in the study population.

In order to perform PR analysis of MRS data, MR scientists nowadays have the choice either to use an existing software package (Matlab, R, Simca, or SpectraClassifier (29) among others) or to collaborate with PR specialists. PR specialists will generally expect to receive a "clean" data matrix with hundreds of data belonging to equally distributed classes (which is not normally the case as we have seen) with any bad-quality data removed, and with each spectrum labelled according to the disease, disease status or outcome variable of interest. In short, a file in which the rows are the cases or measurements and the columns are the variables (peak heights or integrated peak areas). Most PR trials have used the whole "raw" spectrum instead of integrated peaks. This simplifies the trial by removing the need for expert spectroscopist participation before data upload, which as we have seen is a major cause of lost data. More importantly, it eliminates the need for expert assignment and quantification of spectral peaks when the method is being used in clinical practice.

## 3.   ACHIEVEMENTS OF MRS TRIALS

### 3.1.  Proton spectroscopy: Multicenter trials involving pattern recognition (PR)

The INTERPRET database continued to be maintained after the project finished. This fact allowed two other multicenter trials (eTUMOUR(2) and HealthAgents(21)) to use the data already acquired and validated for the further development of classifiers. Those new classifiers were then validated using prospectively acquired spectra. From the PR point of view, the single most important result of these three trials is the demonstration that classifiers for SV data at 1.5T can distinguish robustly between the most common human brain tumor types with >90% accuracy. In this context, robustness means that spectra are still correctly classified if they have been acquired under slightly different acquisition conditions, for example with PRESS or STEAM sequences, or with differences in TR and TE, or instrument manufacturer (8). Due to the time span for acquisition of data (1994-2008), the classifiers can also be considered insensitive to scanner models, operators and even magnetic field up to 3T (30), although the latter was only tested on a small dataset.

The classifiers developed for the most common tumor types have been validated from many points of view (31). For internal validation, classifiers were trained and tested with different datasets (1,8). For temporal validation, classifiers developed with data from one multicenter project were shown to perform well when tested against data acquired during another multicenter project years later (9,32). For external validation, these classifiers and their associated decision-support system were also found to perform well with prospectively acquired data, not necessarily from the same classes used to train the classifiers.  Throughout these tests the results from the classifiers significantly improved radiologists' predictions about tumor type as compared to the use of MRI alone(33-35).  All the acquisition, processing and validation criteria developed in these trials are applicable to future multicenter projects.

The classifiers developed during these trials were only for the most common tumor types. To date, it has not been possible to classify many of the brain tumor types in the World Health Organization (WHO) classification. This is due both to data sparsity and also to the similarity of some of the spectral patterns.  Another issue is that the main multicenter results are SV spectra obtained at 1.5T.

Furthermore, MV processing, co-registration and output, even at 1.5T, are still not yet manufacturer independent. Table 5 gives some recommendations for future trials, from lessons learned with these trials.

**Table 5 near here**

### 3.2. Phosphorus spectroscopy: The Cooperative Group of MRS Applications in Cancer (CoGMAC)

This trial was performed under an NIH-funded international cooperative program. Using MV $^{31}$P MRS in vivo, it demonstrated in newly diagnosed diffuse large B-cell lymphoma (DLBCL) patients that the pretreatment tumor phosphomonoester (PME) values (i.e., the sum of phosphocholine and phosphoethanolamine), when normalized by nucleoside triphosphates (PME/NTP) values identified patient groups with significantly different treatment outcomes (i. e. treatment response at six months and time to treatment failure,TTF) with high effect size (36).

Differences in the mean tumor value of the pre-treatment PME/NTP were determined, using an independent-samples t-test, in DLBCL patients who exhibited either complete (CR) or not complete (NCR) response after six months of treatment (37). The mean PME/NTP was 1.44 (standard deviation, SD=0.43; n=18 subjects) in CR patients and 2.24 (SD=0.57, n=14) in NCR patients, with a highly statistically significant difference ($p < .0001$), and a large effect size ($d = 1.59$). A Fisher's exact test was also conducted to determine association between the pretreatment PME/NTP values and response to treatment at six months (CR *vs.* NCR). There was a statistically significant association ($p < .004$) between the pretreatment PME/NTP and the response to treatment. Overall success rate was 78.1% (sensitivity=57.1%; specificity=94.4%; positive predictive value=88.9%; and negative predictive value=73.9%).

Kaplan-Meier survival analysis was also conducted to compare DLBCL patients set apart into groups with low ($> 2.2$) or high ($\leq 2.2$) PME/NTP value for their likelihood to predict TTF. Participants in the low PME/NTP group had a median TTF of 66.5 (95% confidence interval, CI=32.9 to 100.1) months. This was longer than the high PME/NTP group, which had a median TTF of 6.8 (95% CI=5.2

to 8.1) months. A log rank test was conducted to determine if there were differences in the survival distributions for the different groups. The survival distributions were statistically significantly different ($\chi^2_{(1)} = 14.43$; $p < .0005$).

## RELATED ARTICLES

Emrstm1484, emrstm1496.

## 1. REFERENCES

1. Tate AR, Majos C, Moreno A, Howe FA, Griffiths JR, Arus C. Automated classification of short echo time in in vivo 1H brain tumor spectra: a multicenter study. Magn Reson Med 2003;49(1):29-36.
2. Julià-Sapé M, Lurgi M, Mier M, Estanyol F, Rafael X, Candiota A, Barceló A, García A, Martínez-Bisbal M, Ferrer-Luna R, Moreno-Torres Á, Celda B, Arús C. Strategies for annotation and curation of translational databases: the eTUMOUR project. Database (Oxford) 2012;2012:bas035.
3. Arias-Mendoza F, Payne GS, Zakian KL, Schwarz AJ, Stubbs M, Stoyanova R, Ballon D, Howe FA, Koutcher JA, Leach MO, Griffiths JR, Heerschap A, Glickson JD, Nelson SJ, Evelhoch JL, Charles HC, Brown TR. In vivo 31P MR spectral patterns and reproducibility in cancer patients studied in a multi-institutional trial. NMR Biomed 2006;19(4):504-512.
4. Arias-Mendoza F, Zakian K, Schwartz A, Howe FA, Koutcher JA, Leach MO, Griffiths JR, Heerschap A, Glickson JD, Nelson SJ, Evelhoch JL, Charles HC, Brown TR. Methodological standardization for a multi-institutional in vivo trial of localized 31P MR spectroscopy in human cancer research. In vitro and normal volunteer studies. NMR Biomed 2004;17(6):382-391.
5. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HCW, Lijmer JG. The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration. Annals of Internal Medicine 2003;138(1):W1-12.
6. Lin AP, Tran TT, Ross BD. Impact of evidence-based medicine on magnetic resonance spectroscopy. NMR Biomed 2006;19(4):476-483.
7. Weinreb JC, Blume JD, Coakley FV, Wheeler TM, Cormack JB, Sotto CK, Cho H, Kawashima A, Tempany-Afdhal CM, Macura KJ, Rosen M, Gerst SR, Kurhanewicz J. Prostate cancer: sextant localization at MR imaging and MR spectroscopic imaging before prostatectomy--results of ACRIN prospective multi-institutional clinicopathologic study. Radiology 2009;251(1):122-133.
8. Tate AR, Underwood J, Acosta DM, Julia-Sape M, Majos C, Moreno-Torres A, Howe FA, van der Graaf M, Lefournier V, Murphy MM, Loosemore A, Ladroue C, Wesseling P, Luc Bosson J, Cabanas ME, Simonetti AW, Gajewicz W, Calvar J, Capdevila A, Wilkins PR, Bell BA, Remy C, Heerschap A, Watson D, Griffiths JR, Arus C. Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. NMR Biomed 2006;19(4):411-434.
9. Garcia-Gomez JM, Luts J, Julia-Sape M, Krooshof P, Tortajada S, Robledo JV, Melssen W, Fuster-Garcia E, Olier I, Postma G, Monleon D, Moreno-Torres A, Pujol J, Candiota AP, Martinez-Bisbal MC, Suykens J, Buydens L, Celda B, Van Huffel S, Arus C, Robles M. Multiproject-multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy. Magn Reson Mater Phy 2009;22(1):5-18.
10. Zarino B, Crespi M, Launi M, Casarotti A. A new standardization of semantic verbal fluency test. Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology 2014.
11. van der Graaf M, Julia-Sape M, Howe FA, Ziegler A, Majos C, Moreno-Torres A, Rijpkema M, Acosta D, Opstad KS, van der Meulen YM, Arus C, Heerschap A. MRS quality assessment in a multicentre study on MRS-based classification of brain tumours. NMR Biomed 2008;21(2):148-158.
12. Wright AJ, Arus C, Wijnen JP, Moreno-Torres A, Griffiths JR, Celda B, Howe FA. Automated quality control protocol for MR spectra of brain tumors. Magn Reson Med 2008;59(6):1274-1281.

13. Kuesel AC, Stoyanova R, Aiken NR, Li CW, Szwergold BS, Shaller C, Brown TR. Quantitation of resonances in biological 31P NMR spectra via principal component analysis: potential and limitations. NMR Biomed 1996;9(3):93-104.
14. Stoyanova R, Brown TR. NMR spectral quantitation by principal component analysis. NMR Biomed 2001;14(4):271-277.
15. Stoyanova R, Brown TR. NMR spectral quantitation by principal component analysis. III. A generalized procedure for determination of lineshape variations. J Magn Reson 2002;154(2):163-175.
16. Arias-Mendoza F, Brown TR. *In vivo* measurement of phosphorous markers of disease. Dis Markers 2003-2004;19(2-3):49-68.
17. Vigneron DB, Nelson SJ, Murphy-Boesch J, Kelley DA, Kessler HB, Brown TR, Taylor JS. Chemical shift imaging of human brain: axial, sagittal, and coronal P-31 metabolite images. Radiology 1990;177(3):643-649.
18. Coons SW, Johnson PC, Scheithauer BW, Yates AJ, Pearl DK. Improving diagnostic accuracy and interobserver concordance in the classification and grading of primary gliomas. Cancer 1997;79(7):1381-1393.
19. Julia-Sape M, Acosta D, Mier M, Arus C, Watson D. A multi-centre, web-accessible and quality control-checked database of in vivo MR spectra of brain tumour patients. Magn Reson Mater Phy 2006;19(1):22-33.
20. Scheenen TWJ, Fütterer J, Weiland E, van Hecke P, Lemort M, Zechmann C, Schlemmer H-P, Broome D, Villeirs G, Lu J, Barentsz J, Roell S, Heerschap A. Discriminating Cancer From Noncancer Tissue in the Prostate by 3-Dimensional Proton Magnetic Resonance Spectroscopic Imaging: A Prospective Multicenter Validation Study. Investigative Radiology 2011;46(1):25-33 10.1097/RLI.1090b1013e3181f54081.
21. González-Vélez H, Mier M, Julià-Sapé M, Arvanitis T, García-Gómez J, Robles M, Lewis P, Dasmahapatra S, Dupplaw D, Peet A, Arús C, Celda B, Van Huffel S, Lluch-Ariet M. HealthAgents: distributed multi-agent brain tumor diagnosis and prognosis. Applied Intelligence 2009;30(3):191-202.
22. Zhao Q, Patriotis P, Arias-Mendoza F, Stoyanova R, Brown TR. Interactive chemical shift imaging: a comprehensive software program for data analysis and quantification.; 2006; Warrenton VA, USA.
23. Zhao Q, Patriotis P, **Arias-Mendoza F**, Stoyanova R, Brown TR. Interactive chemical shit imaging: a comprehensive software program for data analysis and quantification. 2006; Asilomar CA, USA.
24. Provencher SW. Estimation of metabolite concentrations from localized in vivo proton NMR spectra. Magn Reson Med 1993;30(6):672-679.
25. Stefan D, Cesare FD, Andrasescu A, Popa E, Lazariev A, Vescovo E, Strbak O, Williams S, Starcuk Z, Cabanas M, Ormondt Dv, Graveron-Demilly D. Quantitation of magnetic resonance spectroscopy signals: the jMRUI software package. Measurement Science and Technology 2009;20(10):104035.
26. Crane JC, Olson MP, Nelson SJ. SIVIC: Open-Source, Standards-Based Software for DICOM MR Spectroscopy Workflows. International journal of biomedical imaging 2013;2013:169526.
27. Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics 2003;19(12):1484-1491.
28. Ostrom QT, Gittleman H, Liao P, Rouse C, Chen Y, Dowling J, Wolinsky Y, Kruchko C, Barnholtz-Sloan J. CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2007–2011. Neuro-Oncology 2014;16(suppl 4):iv1-iv63.
29. Ortega-Martorell S, Olier I, Julia-Sape M, Arus C. SpectraClassifier 1.0: a user friendly, automated MRS-based classifier-development system. BMC Bioinformatics 2010;11.
30. Fuster-Garcia E, Navarro C, Vicente J, Tortajada S, García-Gómez J, Sáez C, Calvar J, Griffiths J, Julià-Sapé M, Howe F, Pujol J, Peet A, Heerschap A, Moreno-Torres À, Martínez-Bisbal M, Martínez-Granados B, Wesseling P, Semmler W, Capellades J, Majós C, Alberich-Bayarri À, Capdevila A, Monleón D, Martí-Bonmatí L, Arús C, Celda B, Robles M. Compatibility between 3T 1H SV-MRS data and automatic brain tumour diagnosis support systems based on databases of 1.5T 1H SV-MRS spectra. Magn Reson Mater Phy 2011:1-8.
31. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.
32. Perez-Ruiz A, Julia-Sape M, Mercadal G, Olier I, Majos C, Arus C. The INTERPRET Decision-Support System version 3.0 for evaluation of Magnetic Resonance Spectroscopy data from

human brain tumours and other abnormal brain masses. BMC Bioinformatics 2010;11(1):581.

33. Julia-Sape M, Coronel I, Majos C, Candiota AP, Serrallonga M, Cos M, Aguilera C, Acebes JJ, Griffiths JR, Arus C. Prospective diagnostic performance evaluation of single-voxel 1H MRS for typing and grading of brain tumours. NMR Biomed 2012;25(4):661-673.

34. Fellows GA, Wright AJ, Sibtain NA, Rich P, Opstad KS, McIntyre DJ, Bell BA, Griffiths JR, Howe FA. Combined use of neuroradiology and 1H-MR spectroscopy may provide an intervention limiting diagnosis of glioblastoma multiforme. J Magn Reson Imaging 2010;32(5):1038-1044.

35. Julià-Sapé M, Majós C, Camins A, Samitier A, Baquero M, Serrallonga M, Doménech S, Grivé E, Howe FA, Opstad K, Calvar J, Aguilera C, Arús C. Multicentre evaluation of the INTERPRET decision support system 2.0 for brain tumour classification. NMR in Biomedicine 2014.

36. Arias-Mendoza F, Payne GS, Zakian K, Stubbs M, O'Connor OA, Mojahed H, Smith MR, Schwarz AJ, Shukla-Dave A, Howe F, Poptani H, Lee SC, Pettengel R, Schuster SJ, Cunningham D, Heerschap A, Glickson JD, Griffiths JR, Koutcher JA, Leach MO, Brown TR. Noninvasive phosphorus magnetic resonance spectroscopic imaging predicts outcome to first-line chemotherapy in newly diagnosed patients with diffuse large B-cell lymphoma. Acad Radiol 2013;20(9):1122-1129.

37. Arias-Mendoza F, Payne GS, Zakian KL, Stubbs M, O'Connor OA, Mojahed H, Smith MR, Schwartz AJ, Shukla-Dave A, Howe FA, Poptani H, Lee S-C, Pettengel R, Schuster SJ, Cunningham D, Heerschap A, Glickson JD, Griffiths JR, Koutcher JA, Leach MO, Brown TR. Phosphorus MR spectroscopy predicts outcome to doxorubicin-based chemotherapy in diffuse large B-cell lymphoma. Acad Radiol (accepted) 2013.

38. Candiota AP, Majos C, Bassols A, Cabanas ME, Acebes JJ, Quintero MR, Arus C. Assignment of the 2.03 ppm resonance in in vivo 1H MRS of human brain tumour cystic fluid: contribution of macromolecules. Magn Reson Mater Phy 2004;17(1):36-46.

## ACKNOWLEDGMENTS

## BIOGRAPHICAL SKETCH

**Margarida Julià-Sapé,** b. 1969, "Licenciada" in Biology 1994, PhD 2006. Joined the "Universitat Autònoma de Barcelona" in 2000 to be part of the team in a multi-center trial on PR of MRS data of brain tumors and, was later involved in other multi-center projects involving MRS with the data manager role. Research interests: applying pattern recognition techniques to MRS data of brain

tumors to improve their diagnosis and follow-up. Other interests: MRS data curation, facilitating the uptake of MRS by clinicians through decision-support.

**Fernando Arias-Mendoza,** b. 1957 is a doctor of Medicine (MD, 1981) and a doctor in Biochemical Sciences (PhD, 1988). Recently joined the Department of Radiology of the University of Pennsylvania Health System. Since 1999 Dr. Arias-Mendoza has been co-principal investigator or principal investigator of research programs focused on the prediction of tumor response to treatment using MRS and advanced MR imaging techniques.

**John Griffiths,** b. 1945. Qualified in medicine and biochemistry. In the early 1980s, his research group pioneered the use of MRS for studies on living tumors, and he has worked since then on MRI and MRS of cancer, both in vivo and ex vivo. He has published more than 300 peer-reviewed articles to date. His recent interests include the metabolomics of cancer.

## TABLES AND CAPTIONS

| |
|---|
| 1. Prospective study. |
| 2. Large patient population (n>50 or power calculations to demonstrate that the size is sufficient). |
| 3. Quantitative analysis. |
| 4. Formal criteria for diagnosis (i.e. ROC). |
| 5. Multiple blinded analyses. |
| 6. Confirmation of MRS results (i.e. biopsy or out- |
| 7. Provide sensitivity and specificity measures. |
| 8. 'Added value' diagnostic enhancement over routine CT, MRI, SPECT or PET. |
| 9. Examine impact on diagnostic thinking (patient management). |
| 10. Examine impact on choice of therapy. |

**Table 1. Common factors for a study on MRS that fulfills EBM criteria.**

| PARAMETER | STEAM (SHORT TE) | PRESS (SHORT TE) | PRESS (LONG TE) |
|---|---|---|---|
| TE | 20 ms (20 – 32 ms) | 30 -32 ms (30-32 ms) | 136 ms (135 – 144 ms) |
| TR | 2000 ms (1600 – 2000 ms) | 2000 ms (1600 – 2000 ms) | 2000 ms (1600 – 2000 ms) |
| Volume | 4 – 8 cm$^3$ | 4 – 8 cm$^3$ | 4 – 8 cm$^3$ |
| Number of averages metabolites | 256 | 192 - 128 | 192 - 128 |
| Number of averages water | 8 to 32 | 8 to 16 | 8 to 16 |
| Number of points | 512 [Philips] 1024 [Siemens] 2048 [GE] | 512 [Philips] 1024 [Siemens] 2048 [GE] | 512 [Philips] 1024 [Siemens] 2048 [GE] |
| Bandwidth | 1000 Hz [Philips] 1000 Hz [Siemens] 2500 Hz [GE] | 1000 Hz [Philips] 1000 Hz [Siemens] 2500 Hz [GE] | 1000 Hz [Philips] 1000 Hz [Siemens] 2500 Hz [GE] |
| Dummy scans | 4 | 4 | 4 |

**Table 2. Consensus INTERPRET acquisition protocols (8) for new data with ranges used for retrospective data accepted into the database. TE and TR ranges used for retrospective data are given in parentheses.**

| ORDER IN WHICH IT WAS PERFORMED | PROCEDURE |
|---|---|
| 1st | Lineshape correction and zero order phasing using water reference with the Klose method |
| 2nd | 0.8 Hz exponential line broadening |
| 3rd | Processing by FFT |
| 4th | Water removal by HLSVD: 5 components removed within $\pm0.37$ ppm of water resonance |
| 5th | Residual water suppression: points at 4.2 to 5.1 ppm set to zero |
| 6th | Linear interpolation to 512 points over 1000 Hz of Siemens and Philips data |
| 7th | Spectrum alignment: maximum of choline peak shifted to 3.21ppm |
| 8th | Normalization of spectrum to Euclidian norm of peak heights |

**Table 3. Consensus data processing into the INTERPRET canonical format (8) for spectrum display and analysis.**

| Trial | DATA TYPE | PROCESS | PARAMETERS | ACCEPT IF |
|---|---|---|---|---|
| INTERPRET and eTUMOUR | SV | Automatic | Water linewidth (WLW): half-height linewidth of the water peak of the unsuppressed water spectrum | WLW < 8Hz |
| INTERPRET and eTUMOUR | SV | Automatic | Signal-to-noise ratio (SNR). Signal = Maximum metabolite signal in range 0 – 3.4ppm; Noise = standard deviation noise in range 9 – 11ppm. SNR = Signal/Noise | SNR > 10 |
| INTERPRET | SV | Expert-based | Possible artefacts that cause rejection: high scalp lipids; poor phasing; large baseline artefacts; metabolite peaks of suspect origin | 2 experts out of 3 agree |
| eTUMOUR | SV | Expert-based | Excellent: Narrow lines, high signal-to-noise, flat baseline and no artefacts. Good: only 2 of the above and no artefacts. Poor: only one of the above and no artefacts. Unacceptable: Artefacts and /or outside criteria WLW<8 Hz and SNR>10 | 2 experts out of 3 agree |
| eTUMOUR | MV | Expert-based | Excellent: At least 2 of the following: narrow lines, high SNR, flat baseline; and no major artefacts in >80% of voxels Good: 50% - 80% of voxels are excellent Poor: < 50% of voxels are excellent Unacceptable: Less than a 4x4, 5x3 or 3x5 voxel region are usable Additional reasons for bad quality: Poor SNR, Poor Water Suppression, Scalp Artefacts, and too Few Usable Voxels. | 2 experts out of 3 agree |
| eTUMOUR | HRMAS | Expert-based | Grading: Excellent, Good, Poor, And Unacceptable. Reasons for bad quality: Contamination from Ethanol, Contamination from Acetone, Contamination from Mannitol, Contamination from Other Substances, Low SNR, Low Metabolites, Split Peaks, Negative Peaks, Truncation, Insufficient Water Suppression, Rotation Speed Side Bands, Unphaseable, Poor resolution | 2 experts out of 3 agree |

**Table 4. Consensus QC of SV data in INTERPRET (8) and SV, MV and HRMAS in eTUMOUR (2).**

| **Reccomendations** | |
|---|---|
| 1. Build from previous experience. | Work done at lower magnetic fields can help to define more precise clinical questions |
| | Successful strategies used in the past should be applied again and those that do not work should be eliminated |
| 2. Standardize. | Protocols for: Acquisition, processing, quality control and data exchange |
| | Define the minimum set of information needed for quality control and for reporting an experiment |
| 3. Automate spectral processing. | Saves expert time |
| | Avoids expert bias. |
| 4. Share data. | The same data can be analyzed in different ways |
| | Favors reproducibility of research results |
| 5. Keep logistics simple. | Avoid centralizing procedures on one single expert, especially for quality control |
| | Avoid too many different protocols but accept a certain degree of flexibility as long as it provides a reasonable number of good quality data |

**Table 5. Summary of reccomendations derived from lessons learned in past pattern recognition trials.**
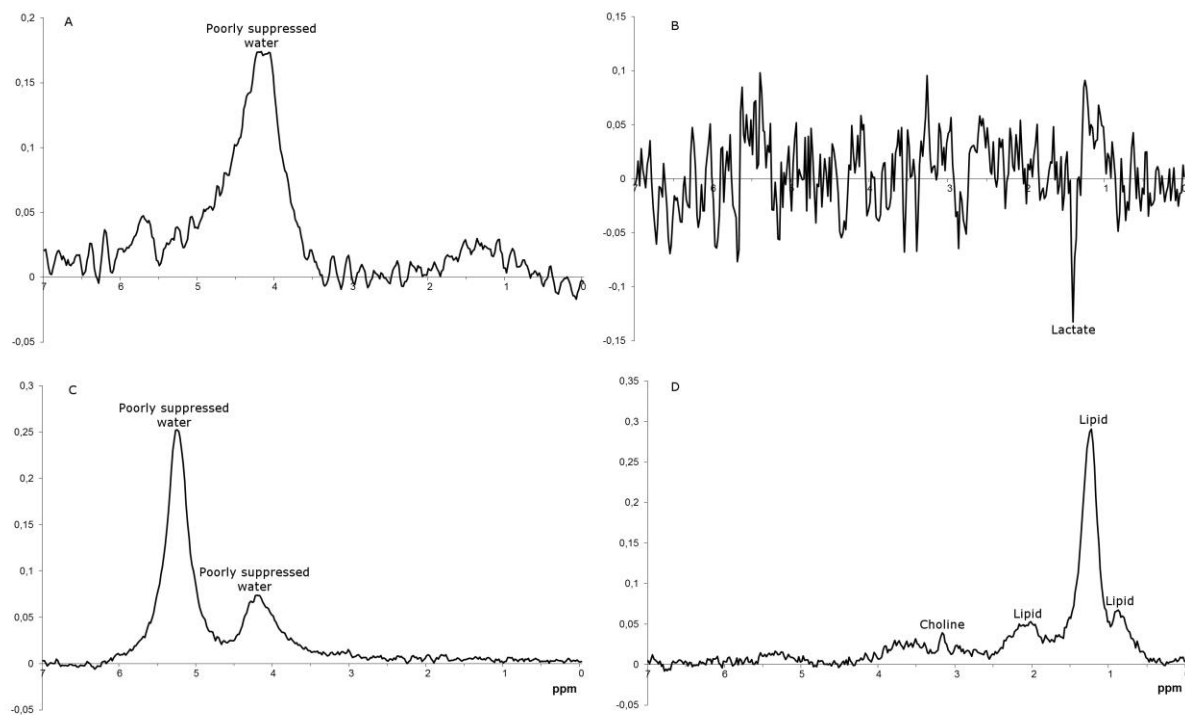
**FIGURES AND CAPTIONS**



**Figure 1. SV $^1$H brain tumor spectra at 1.5T from the INTERPRET database, after being processed with the latest INTERPRET pipeline ((32) except that the residual water region has not been set to zero). The Y axes are in arbitrary units after Euclidean norm. Spectra were acquired with the INTERPRET protocol (8). Voxel sizes for these cases were between 1 and 3.375 ml. Identification number (ID). The spectra have been reprocessed for this figure and only 0-7 ppm are displayed for visualzation purposes.**

**A): ID=I0301/svs000/idf0001: Bad SNR and bad WLW (see Table 4 for definitions of SNR and WLW), STEAM, TE=20 ms, SNR=5.61, WLW=29.22 Hz. Acquired in 1997.**

**B) ID I0100/svs000/idf0002: Bad SNR and good WLW: STEAM, TE=135 ms, SNR=5.92, WLW=2.75 Hz. Acquired in 1997.**

**C) ID=I0230/svs000/idf0000: Good SNR and bad WLW, STEAM, TE=135 ms, SNR=8.38, WLW=21.59 Hz. Acquired in 1994.**

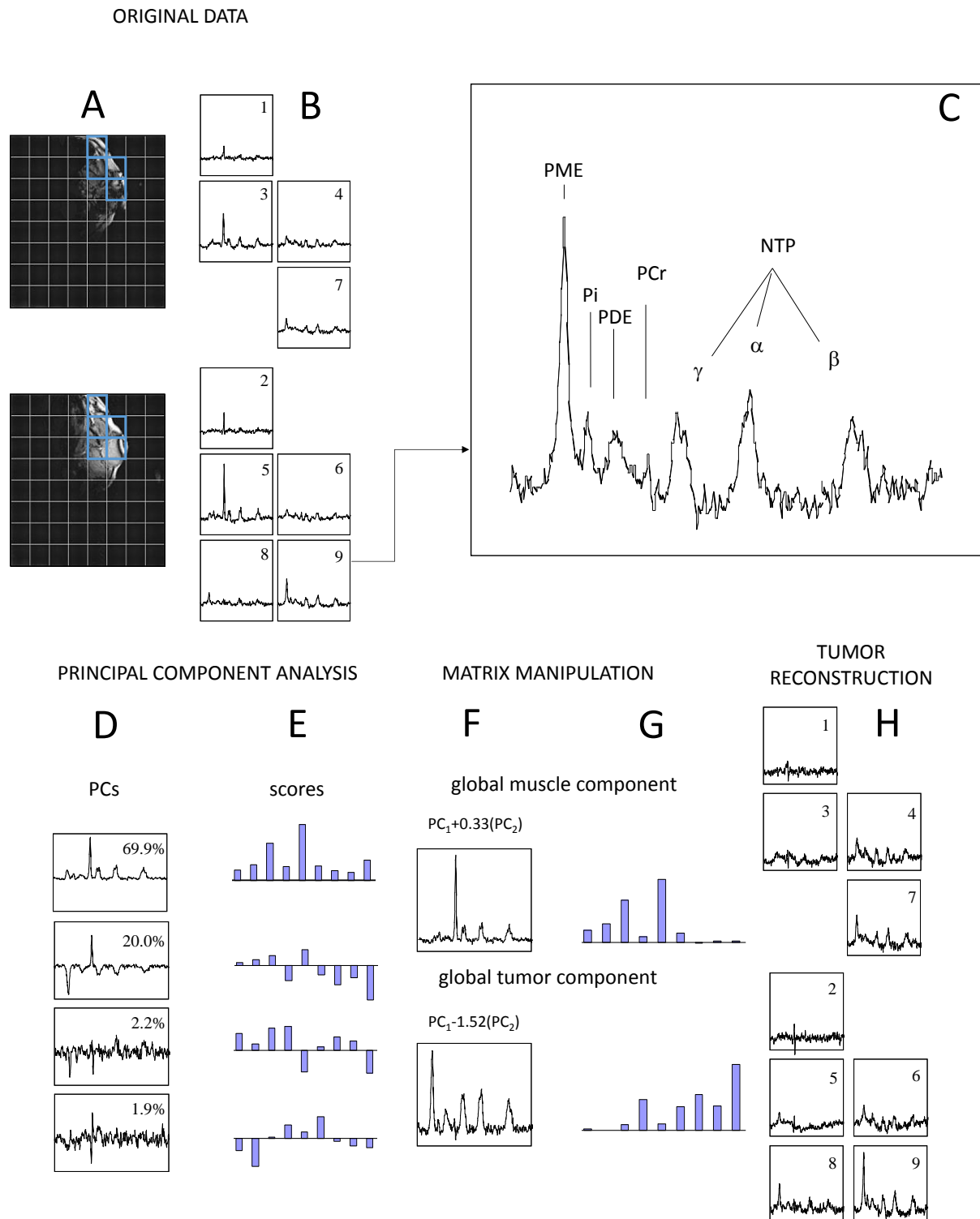**D) ID I1161/svs000/idf0000: Good SNR and good WLW, STEAM, TE=30 ms, SNR=93.32, WLW=2.29 Hz. Acquired in 2001.**

Figure 2. **$^{31}$P MR spectral set of an individual diffuse large B-cell lymphoma patient. Column A under "Original Data" show low-resolution MR images (obtained with a surface coil) overlaid with the grid of volumes of the spectral dataset. The images also show 9 highlighted volumes matching with the lymphoma mass. Column B shows the spectra corresponding to the highlighted volumes in the images. Panel C expands spectrum 9 of column B, which has the largest tumor component, to show its quality and peak assignments. PME, phosphomonoesters including phosphocholine and phosphoethanolamine; Pi, inorganic phosphate; PDE, diphosphoester region where also phospholipid resonate; PCr, phosphocreatine, small in this spectrum but the largest signal in muscle (i.e., spectra 2, 3, and 5 in column B); NTP the γ, α, and β signals of nucleoside triphosphates. Columns D and E under "Principal Component Analysis" show the principal components (PC's) and associated scores obtained by the analysis. Column**

D shows that 89.9% of the variance is included in the first two PC's. The addition of these first two PCs yielded a clean "global muscle component" while subtraction yielded a clean "tumor component" as shown in the columns F and G under "Matrix Manipulation". Further, matrix inversion and multiplication to eliminate the muscle component allowed analysis of the percent contribution of tumor in each spectrum (column H, "Tumor Reconstruction". Compare spectrum 5 in columns B and H where a strong muscle component (column B) obscured a small but important tumor contribution in the spectrum (column H). Similar results were obtained by analyzing only the nine spectra shown in the figure, a subset of 30 spectra around the tumor region, or the total dataset of 512 spectra.
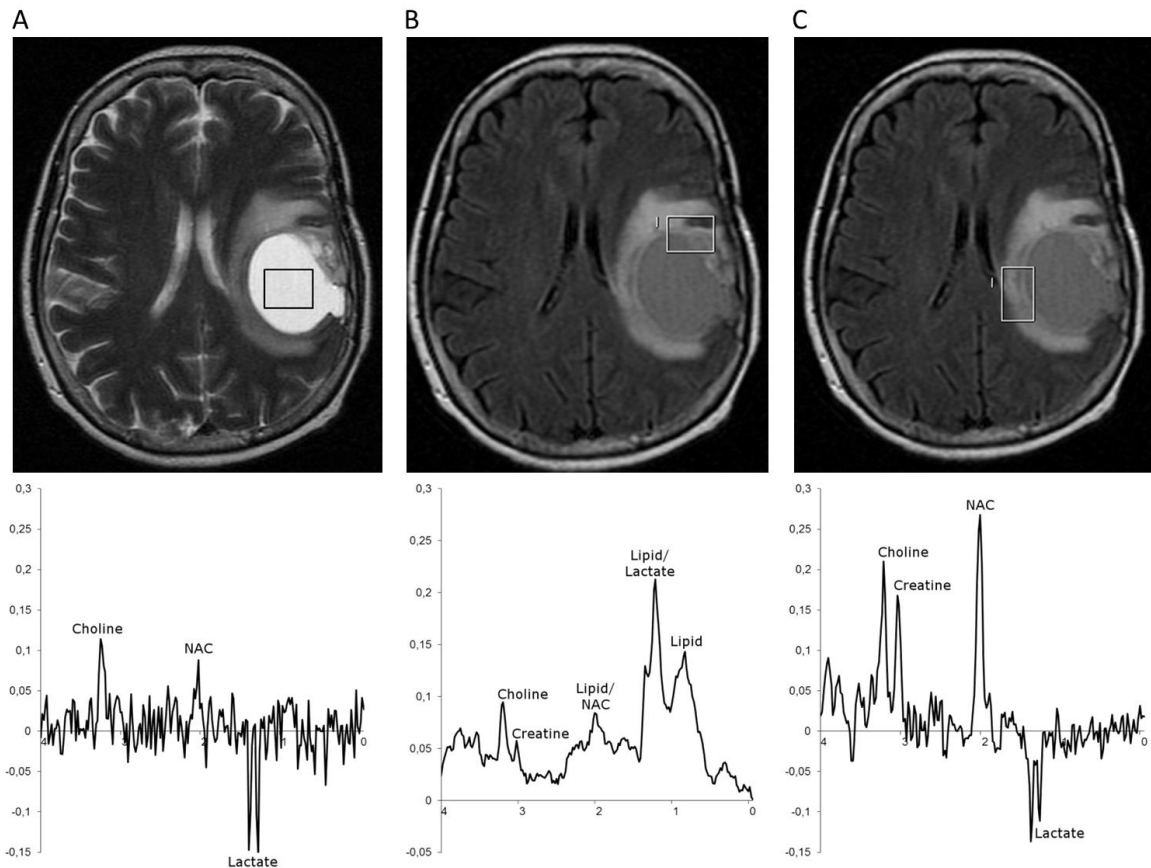


Figure 3. : Anaplastic astrocytoma (case I0456 from the INTERPRET project (8)), processed with the automated software developed. Note that it is not only important to know the type of disease the spectrum corresponds to but also the importance of a proper annotation of the location and other metadata: the spectrum in A is not comparable to the one on B in terms of pattern, although they both belong to the same patient and pathology. A was recorded well into the cystic liquid while C was acquired in the tumor margins. B was acquired in the margins of the tumor but at a different echo time (short).

A) Top, reference image with voxel location inside cystic lesion. PRESS, TE=136 ms, STR=2000, SNR=12.17, WLW=2.52 Hz.

B) Top, reference image with voxel location on tumour margin. PRESS, TE=30 ms, STR=2000, SNR=59.99, WLW=3.43 Hz

C) Top, reference image with voxel location on tumor margin. PRESS, TE=136 ms, STR=2000, SNR=16.64, WLW=2.06 Hz.

SNR and WLW calculated as in Figure 1. Y axis in arbitrary units as in Figure 1. X axis, in ppm. NAC, N-acetylated compounds (38).