Sáez Silvestre, C.; Martí-Bonmatí, L.; Alberich Bayarri, A.; Robles Viejo, M.; García Gómez, JM. (2014). Randomized pilot study and qualitative evaluation of a clinical decision support system for brain tumour diagnosis based on SV 1H MRS: Evaluation as an additional information procedure for novice radiologists. Computers in Biology and Medicine. 45:26-33. doi:10.1016/j.compbiomed.2013.11.009.

Randomized pilot study and qualitative evaluation of a clinical decision support system for brain tumour diagnosis based on SV 1H MRS: Evaluation as an additional information procedure for novice radiologists

Carlos SÁEZ[a1] , Luis MARTÍ-BONMATÍ[b,c], Ángel ALBERICH-BAYARRI[b], Montserrat ROBLES[a] and Juan M GARCÍA-GÓMEZ[a]

[a] *Grupo de Informática Biomédica (IBIME), Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València Camino de Vera s/n, 46022 Valéncia, Spain*

[b] *Department of Radiology, Hospital Quirón Valencia, Valencia, Spain*

[c] *Radiology, Department of Medicine, Universidad de Valencia, Spain*

Abstract

The results of a randomized pilot study and qualitative evaluation of the clinical decision support system Curiam BT are reported. We evaluated the system's feasibility and potential value as a radiological information procedure complementary to magnetic resonance (MR) imaging to assist novice radiologists in diagnosing brain tumours using MR spectroscopy (1.5 and 3.0 Tesla). Fifty-five cases were analysed at three hospitals according to four non-exclusive diagnostic questions. Our results show that Curiam BT improved the diagnostic accuracy in all the four questions. Additionally, we discuss the findings of the users' feedback about the system, and the further work to optimize it for real environments and to conduct a large clinical trial.

Keywords. randomized pilot study, clinical decision support systems, brain tumours, radiological information procedure, qualitative evaluation

Introduction

Conventional magnetic resonance (MR) images provide highly detailed morphological and microstructural information, and are fundamental in the diagnosis and grading of brain tumours. Although the development of contrast-enhanced and diffusion-weighted MR imaging have greatly improved the diagnostic accuracy of MR imaging, the accurate characterization of tumours remains problematic. During the last decade, magnetic resonance spectroscopy (MRS) has demonstrated its capability to complement MR imaging for initial diagnosis exam of brain masses [2], based on the modification of the metabolic information of different types of brain tumors [3].

Several studies have applied pattern recognition techniques to classify brain tumours based on Proton ^{1}H MRS signals [4-7]. Nevertheless, the difficulty of interpreting the signal is a major impediment to the introduction of such technology in routine clinical practice [2][4][19]. Horská and Baker [19] suggested that automated procedures to analyze MRS and display its results are needed to overcome this issue. For this reason, translational research has focused its attention in developing and evaluating clinical decision support systems (CDSS) based on ^{1}H MRS to help radiologists in the diagnosis of brain tumors [4][8-10]. Additionally, a CDSS for brain tumour diagnosis may be of special interest to novice radiologists, where the lack of clinical experience on large number of real cases of specific tumor types, offers an optimal opportunity for the use of a CDSS [15]. In fact, the evaluation of CDSSs with novice clinicians has been largely addressed in the literature [16][17].

In this work, we present the results of a randomized pilot study to evaluate the feasibility and to define the potential value for clinical practice of Curiam BT, a CDSS for brain tumour diagnosis based on ^{1}H MRS (an earlier abstract was presented in [24]). The evaluation was carried out based on a prospective parallel-randomized pilot trial, in which resident and expert radiologists from three hospitals were

[1] Corresponding author email: carsaesi@ibime.upv.es, Phone: (+34) 963 877 000 Ext. 75278, Fax: (+34) 963 877 279

involved in both quantitative and qualitative assessments. To the best of our knowledge this is the first multi-center randomised pilot study of a CDSS for brain tumour diagnosis using both 1.5 and 3.0 Tesla (T) cases, where we provide evidence to support the feasibility of large scale multi-centre trials in this area [19].

Materials and methods

The design of the pilot study to evaluate Curiam BT with novice radiologists was twofold. First, we carried out a quantitative evaluation based on a prospective parallel-randomized trial. Second, we completed the study with a qualitative evaluation based on the Technology Acceptance Model (TAM) methodology [14], a feedback questionnaire and personal interviews about the clinical use and value of the system.

Curiam BT

Curiam BT is a CDSS for brain tumour diagnosis based on the analysis of 1.5 and 3.0*T*-Single Voxel (SV) ^{1}H MRS data. Curiam BT is a specialization of the generic framework for CDSS Curiam [20] (Figure 1). Curiam provides the generic user interface and logical software components as a basis to build CDSS for specific clinical problems. In addition, it uses the generic classification framework published by the authors in [12], which permits easily including new predictive models based on different pattern recognition or artificial intelligence methods.

From a user-centered approach, the objective of the generic Curiam framework was simple: to offer a single-purpose tool to clinicians who may require support in its decisions. We only focused to one role: clinicians; and to one high-level use case: obtain support for a clinical case based on one or more questions to solve. Thus, specific requirements and use cases are defined when building a specialized version of Curiam for a specific clinical problem.

Based on the Curiam framework several CDSS have been developed for different clinical domains: soft tissue tumour classification and grading (Curiam STT) [20], postpartum depression prediction (Curiam PPD) [20][22], classification of paediatric brain tumours (Curiam BT-Kids) [21], and the CDSS evaluated here, Curiam BT [8].

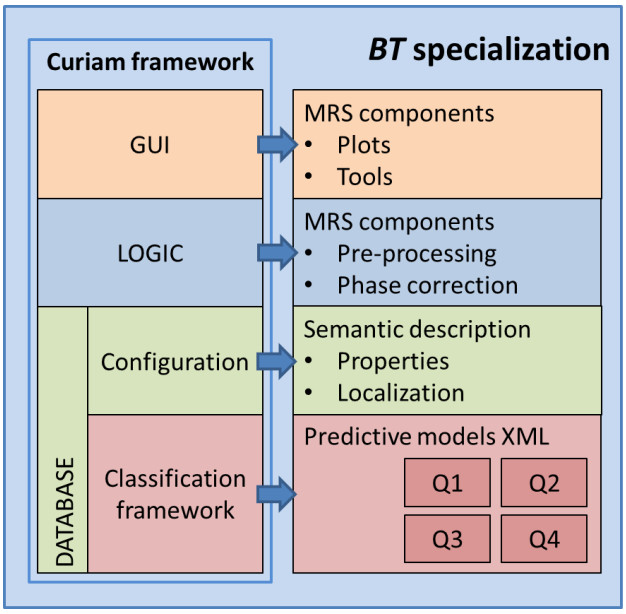

Figure 1. Curiam BT specialization based on the generic Curiam CDSS framework. Following the application program interface (API) of Curiam, only components to manage MRS, a semantic description, and the brain tumour predictive models were required.

Based on the knowledge acquired during the European projects eTumour [1] and HealthAgents [18] we defined several of the requirements and developed Curiam BT. Figure 2 shows the more fine-grained use case of Curiam BT**.** The input of Curiam BT is a raw short time echo (STE) signal (~20ms) alone or in combination with a raw long time echo (LTE) signal (~136ms) acquired with the acquisition protocol described in [5]. The automatic MRS preprocessing pipeline carried out by Curiam BT is based on jMRUI [11] and DMS [4] and is fully described in [6]. An additional zero-order and first-order phase manual correction, especially useful for 3.0*T* signals [7], can be also performed by Curiam BT.

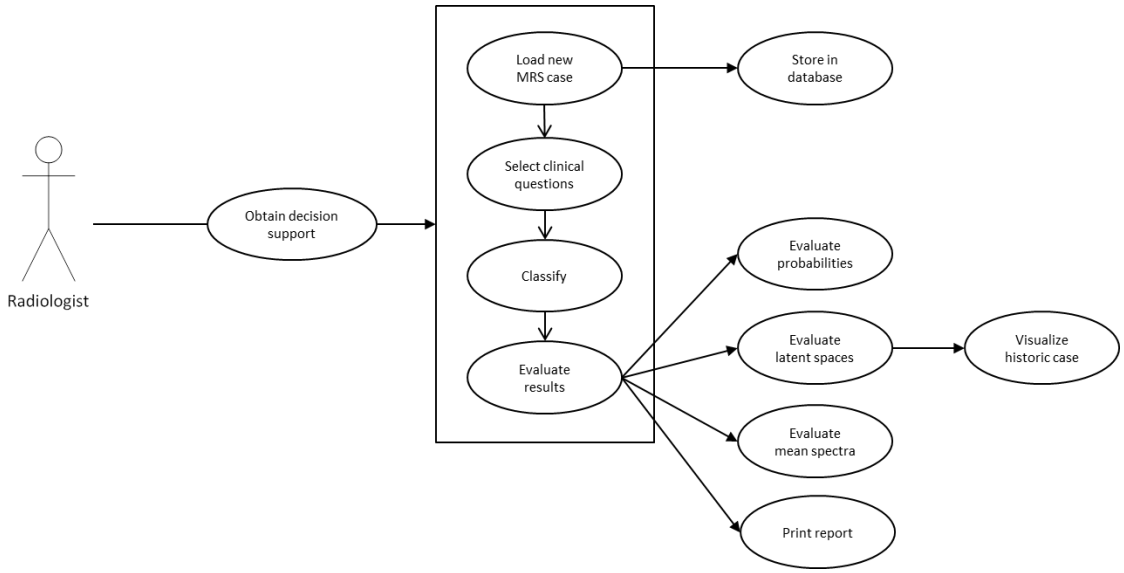

Figure 2. Use case of Curiam BT.

The CDSS used in the pilot study included the four predictive models listed in Table 1. Models M1 and M2 are the STE and combined time echo classifiers for discriminating Aggressive (Glioblastoma and Metastasis), Meningioma and Low Grade Glial (Oligoastrocytoma, Oligodendroglioma and Astrocytoma Grade II) tumours developed and evaluated in [5]. Model M3 was developed and evaluated in [12] to discriminate High Grade Tumours (grade III and IV) and Low Grade Tumours (grade I and II). The three models are based on Fisher Linear Discriminant Analysis (LDA) over the variable space defined in [6] (Peak Integration of metabolites in STE MRS).

Finally, a fourth new classifier, M4, was included in the CDSS to discriminate Meningioma from Non-Meningioma tumours. Non-Meningioma consists of Aggressive and Low Grade Glial tumours as defined in models M1 and M2. This model is also based on Fisher LDA and Peak Integration. The model was trained with 195 cases (from the INTERPRET [4] database: 58 Meningioma, 159 Non-Meningioma), and tested with an independent set of 177 cases (from the eTumour [1] database: 16 Meningioma, and 161 Non-Meningioma). The accuracy (ACC) of the model was 91%.

Predictive model	Discriminated classes	Spectra	ACC (%)
M1	Aggressive vs. Meningioma vs. Low Grade Glial	STE	88
M2	Aggressive vs. Meningioma vs. Low Grade Glial	STE + LTE	92
M3	High Grade Tumour vs. Low Grade Tumour	STE	83
M4	Meningioma vs. Non-Meningioma	STE	91

Table 1. List of predictive models included in the evaluated version of Curiam BT. The established classes or groups are based on the World Health Organization (WHO) classification of tumours of the central nervous system [13]. Aggressive tumours include Metastasis and Glioblastoma; Low Grade Glial includes Oligoastrocytoma, Oligodendroglioma and Astrocytoma Grade II; Low Grade Tumour includes Grade I and II tumours, and High Grade Tumour includes Grade III and IV. STE: Short TE, LTE: Long TE.

The main information provided by Curiam BT consists of the four predictions obtained from the predictive models for the case under study (see Figure 3), where the posterior probabilities calculated by each model are displayed. Additionally, the system provides latent space projections to compare the case under study with similar reference cases used to train the models, labelled according to their histopathological diagnoses. Moreover, the user is able to compare the current spectra with the mean

spectra and standard deviation of the diagnostic classes or groups distinguished by each model. Additionally, the system permits generating PDF or MS-Word reports of the classification results (an example of a report is included in Supplementary Material 3). As additional information, Figure 4 shows the manual phase correction dialog of Curiam BT.

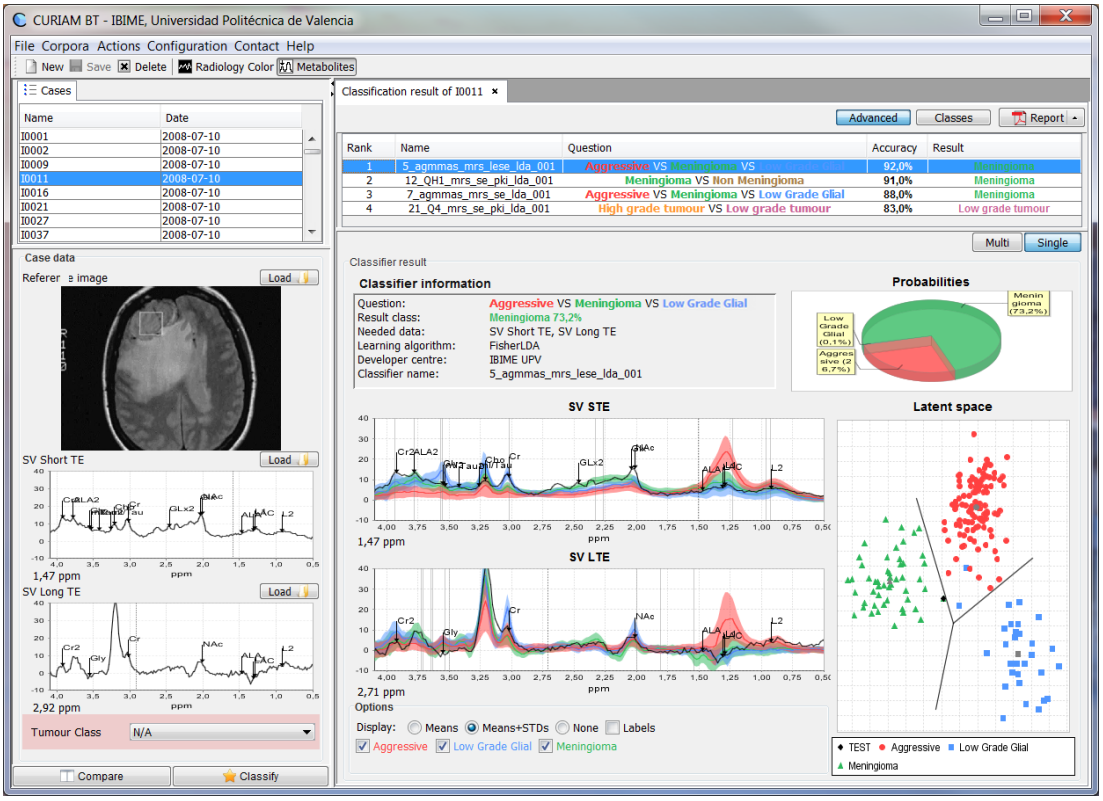

Figure 3. Screenshot of Curiam BT. On the left, the user can select a case from the database or insert a new one. When inserting a new case the user can load a reference image and the corresponding STE and/or LTE spectra. The user can classify the selected case using all or a selection of the included predictive models. On the right the results of a classification are shown. The list at the top shows the resultant diagnosis of each predictive model. Below are shown the posterior probabilities, LDA latent space projection and mean spectra comparison of the selected model. All MRS plots can show the identifiers of the most important metabolites by switching the corresponding button in the toolbar; in addition, they are zoomable, show the PPM index of the cursor position, and their colors can be inverted. The LDA latent space permits visualizing cases similar to the case in-hand to observe their validated diagnoses and apply similar treatments. By the report button, users can generate PDF or MS-Word reports containing the classification results.

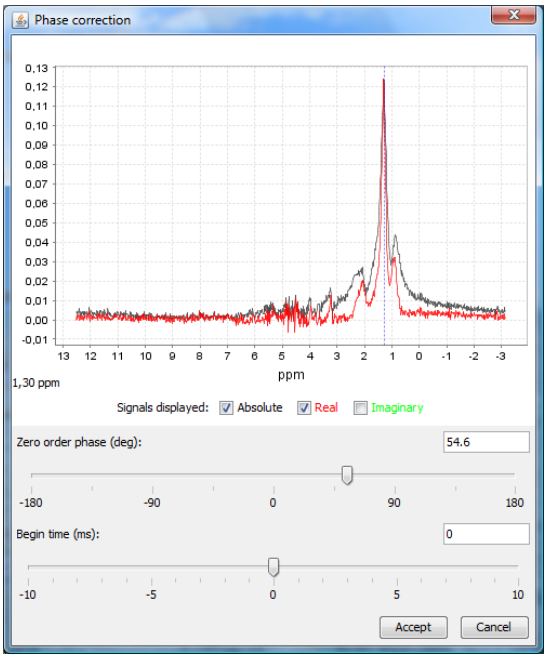

Figure 4. Manual phase correction dialog of Curiam BT. After loading RAW MRS files, the user can manually perform a zero-order and first-order phase correction by means of moving the corresponding slidebar or setting a specific value in its associated text boxes. Also, the user can select whether to show the absolute, real or imaginary signals. This manual correction results especially useful in 3.0*T* spectra[7].

Quantitative evaluation

The quantitative evaluation of the pilot study consisted of a prospective parallel randomized trial. The research aim was to determine the effect on the diagnostic accuracy of novice clinicians when using Curiam BT as an additional radiological information procedure. Consequently, cases were randomized –using simple randomization– into two parallel groups: control and experiment. Cases in the control group were diagnosed using common radiology tests, comprising conventional MR imaging together with diffusion and perfusion weighted images. In the experiment group, Curiam BT was used as an additional radiological information procedure in their common workflow. The system was installed in the radiology services with the supervision of the IT staff of the hospitals.

The evaluation was carried out by three novice radiologists (radiologists in their last year of residence) from three clinical centres, namely the Hospital Quirón Valencia, Hospital Universitario de La Ribera (Alzira) and Hospital Universitario Dr. Peset (Valencia). The participants were asked to diagnose each case according to four non-exclusive diagnostic questions: Q1) High Grade vs. Low Grade vs. N/A; Q2) Grade I vs. II vs. III vs. IV vs. N/A; Q3) Aggressive Tumour vs. Low Grade Glial vs. Meningioma vs. Other; and Q4) Radiologic diagnosis based on WHO [13] –where N/A means 'not applicable'. In the experiment group, the CDSS allowed participants to classify MRS using the four simultaneous predictive models listed in Table 1. Users were free in their interpretation of CDSS results. Before the evaluation, all participants passed a formative period on the use of Curiam BT. On the other hand, the reference radiologic diagnoses were provided by two expert radiologists. The cases were distributed among them, who used all the available information of the case to provide their diagnoses. The expert radiologists also provided their advice and feedback in the development of the study.

The evaluation metrics were defined according to a multiple-class setting. Then, for each class in each question we calculated its recall (REC) and precision (PREC). Note that in a classical two-class setting comprising positive and negative classes, sensitivity (SEN) and specificity (SPE) correspond to REC of positive and REC of negative respectively. Thus, for each question we calculated the ACC (Equation 1) and the macro-averaged recall (REC_M). The REC_M provides an average per-class effectiveness to identify class labels [27]. Equation 2 shows the REC_M formula, calculated as the arithmetic mean of the recalls of each class.

$$ACC = \frac{\sum_{c \in C} TP_c}{N}$$

Equation 1. Formula of accuracy in a multiple-class setting. TP_c corresponds to the true positive answers of the diagnostic class or group c pertaining to the set C of classes. Note that there is no negative class, as each addend considers c as the positive class. N corresponds to the total number of cases.

$$REC_M = \frac{\sum_{c \in C} \frac{TP_c}{TP_c + FN_c}}{|C|}$$

Equation 2. Formula of macro-averaged recall [27] in a multiple-class setting. TP_c and FN_c correspond, respectively, to the true positive and false negative answers of the diagnostic class or group c pertaining to the set C of classes. N corresponds to the total number of cases.

Patients

The study was approved by the ethics committee of the participating hospitals. Data protection was guaranteed as any information that might disclose the identity of the patients was anonymized.

Inclusion criteria consisted of adult patients who were referred to the Radiology Department with the suspicion of a brain tumour, having signed the informed consent. Exclusion criteria consisted of patients' incompatibility to have an MR examination and those with more than one brain tumour or inflammatory-demyelinating lesions.

From an initial set of 69 eligible cases, a total of 55 cases, 27 in experiment and 28 in control, were included in the final analysis after discarding those that met exclusion criteria or were discontinued because no reference diagnosis was provided (due to time limitations of the project). Cases were assigned to the corresponding group at each hospital. STE and LTE MRS were acquired in all the analysed cases, where 27 cases were acquired at 1.5*T* and 28 at 3.0*T* (randomly distributed in control and experiment). The radiological and histopathological diagnoses of the included cases were unknown to the participants at the moment of the study. Figure 5 shows the CONSORT flow diagram of the study.

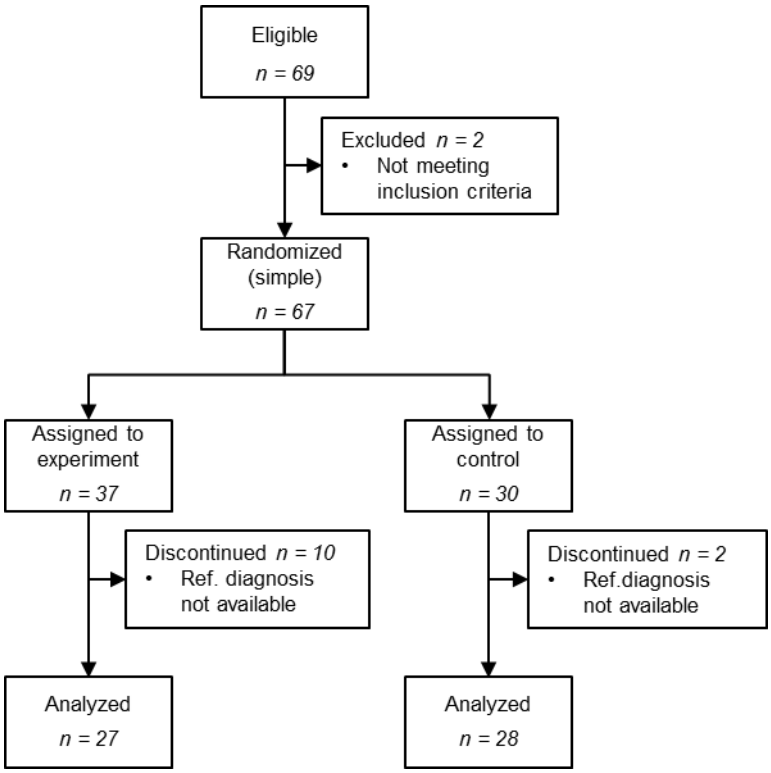

Figure 5. CONSORT flow diagram of the study. From an initial set of 69 eligible cases, the final analysis consisted of 27 cases in the experiment group and 28 in the control group, after discarding those that met exclusion criteria or were discontinued because no reference diagnosis was provided.

Table 2 shows the distribution of diagnostic classes on each clinical question to solve based on the final diagnosis for all the cases included in the pilot study. It can be observed that most cases correspond to astrocytic tumours and to non-tumour cases. Those cases labelled as N/A in Q4, correspond to non-tumour cases. Also, metastatic tumours are labelled as N/A in Q2 as no grade is associated to such tumour type in the WHO classification.

Q4	Q3	Q2	Q1	n for Q2, Q1
Astrocytic tumour (n=27)	Aggressive (n=14)	III	HG	2
		IV	HG	12
	Low Grade Glial (n=13)	I	LG	8
		II	LG	5
Oligodendrogial tumour (n=2)	Low Grade Glial	II	LG	2
Meningioma (n=3)	Meningioma	I	LG	2
		II	LG	1
Cranial and paraspinal nerves tumour (n=1)	Other	I	LG	1
Metastastic tumour (n=2)	Aggressive	N/A	HG	2
Neuronal and mixed neuronal-glial tumours (n=2)	Low Grade Glial	I	LG	2
N/A (n=18)	Other	N/A	N/A	18

Table 2. Frequency of analyzed cases grouped by the clinical questions to solve. N/A means 'not applicable'. The last column shows the joined frequency of cases grouped by the clinical questions.

Qualitative evaluation

The evaluation of user acceptance was carried out by means of a questionnaire, based on the Technology Acceptance Model (TAM) methodology [14]. The TAM provides two constructs or groups of questions. The first evaluates the users' perspective on usefulness and the second evaluates the ease of use of the application. Each construct is composed by six Likert-scale questions ranging from "Totally

disagree" to "Totally agree". The internal consistency of both constructs is demonstrated in [14]. However, from past experience we noted a central tendency bias in users' answers. Thus, to alleviate this bias, in this study we decided to reduce the size of the Likert-scale answers to five items. Additionally, to reduce a possible social desirability bias, questionnaires were answered anonymously by means of an online tool.

The TAM questionnaire was complemented with two questions regarding the assessment for future improvements of the CDSS, both related to its usefulness and to its ease of use: 1) What would you improve to make Curiam BT more useful in your work?, 2) What would you improve to make Curiam BT easier to use?. The full questionnaire is shown in Supplementary Material 1.

To be confident in their answers, the three novice radiologists involved in the quantitative evaluation and two expert radiologists answered the questionnaire after having acquired sufficient experience in the use of Curiam BT during the evaluation period. Finally, the qualitative study was completed with the feedback obtained from personal interviews with the involved radiologists during the development of the project.

Results

Quantitative evaluation

Table 3 shows the ACC and REC_M for each of the evaluated questions. In the four questions, the experiment group provided a higher diagnostic accuracy and REC_M. We may highlight the improvement of the experiment group when answering Q2. This improvement is especially observed when measuring REC_M because of the imbalance of the diagnoses (the REC_M can signal the presence of over or underestimated accuracy results due to class imbalance). On the other hand, it was not possible to calculate the REC_M measurement for Q4 due to the fact that some of the class labels present in the definitive expert diagnoses were not present in the control and experiment groups.

	ACC (95% CI)		REC_M	
	Control	Experiment	Control	Experiment
Q1) HG vs. LG vs. N/A	.64 (.46, .79)	**.70 (.51, .84)**	.64	**.72**
Q2) Grade I vs. II vs. III vs. IV vs. N/A	.46 (.30, .64)	**.59 (.41, .76)**	.46	**.68**
Q3) AGG vs. LGG vs. MEN vs. Other	.68 (.49, .82)	**.74 (.55, .87)**	.75	**.83**
Q4) WHO radiologic diagnosis	.64 (.46, .79)	**.67 (.47, .81)**	-	-

Table 3. Estimated accuracy (ACC) and macro-averaged recall (REC_M) for each question. In the ACC columns the 95% confidence interval (CI) is reported (calculated based on the modified Wald method [23]). For this sample size confidence intervals overlap in control and experiment groups and no significant difference is achieved.

We used these results as a baseline for the estimation of the total number of cases required to obtain statistical significant difference (α=0.05) in a larger RCT for the discrimination of tumour grades (Q2). The estimation yielded a number of 117 cases with the current statistical power for Q2 of 0.50, or 237 cases with a statistical power of 0.80.

The calculated REC and PREC for each of the classes on each question is reported in the Supplementary Material 2. These results show that in general there are no situations with high accuracy and low precision. However, note that the distribution of the number of cases among the different classes (as it can be observed in Table 2) causes some of these to have few individuals in some of the questions. It may increase the variability of the measurements in some questions.

Qualitative evaluation

The evaluation of the questionnaires was carried out by five clinical users. According to the results the usefulness of Curiam BT for their task was valued with an average of 3.58 ($\pm$0.20), and its ease of use was valued with 4.29 ($\pm$0.29).

Regarding the additional question for suggestions to improve the usefulness of the system, radiologists suggested adding additional predictive models to the CDSS to increase its value in routine

clinical practice. Regarding the diagnostic/pre-operative radiological advice, it was suggested that the system should indicate whether we deal with a tumour or not: 1) Tumour vs. Non Tumour, 2) Is it an infection (e.g., abscess)?, 3) Is it a lesion (e.g., demyelinating lesion)?; in case of a tumour, which is its type: 4) Glioblastoma vs. Metastasis, 5) Classification of subtypes of Low Grade Gliomas; and which is its grade: 6) Grade I vs. II, 7) Grade III vs. IV, 8) Grading of meningioma. Regarding the post-operative assessment, it was proposed as a very useful question for patient's follow-up the discrimination between 9) Tumor recurrence vs. Radionecrosis.

For an additional utility of the CDSS, it was also suggested that the system would be useful for comparing SV MRS acquisitions of the affected tissue with their contralateral normal region of the brain. Additionally, radiologists suggested incorporating functionality to subtract signal from necrotic tissue from the MR spectra and the analysis of multi-voxel protocols to allow a better comparison of those cases with problems in the positioning of the voxel during the acquisition. Finally, it was noted that in some bad quality cases (e.g., due to a bad positioning of the voxel) some of the predictive models included in the CDSS were inconsistent among their answers. These inconsistences may highlight problems with the case or its quality; however, they should be specifically managed by the system to facilitate its understanding by the users. Regarding the additional question on suggestions to improve the ease of use, two of the novice radiologists suggested incorporating a full-automatic method to correct the zero- and first-order phase.

On the other hand, we also describe the findings from the continuous feedback collected during the development of the project.

Regarding the usefulness of Curiam BT expert radiologists stated that, in general terms, novice radiologists would be more influenced by the CDSS than them, in congruence with the premise of our evaluation with such participants. Expert radiologists suggested that the CDSS could be a successful tool for the training of novice radiologists as it introduces MRS and makes them reason about the diagnoses while permitting the possibility to compare the current case with similar past cases (in a case-based learning platform). During these interviews, users expressed that the CDSS is particularly useful for performing their task as a diagnostic confirmation tool for routine cases, which gives them more confidence on their diagnoses. It was also reported to be useful as a tool for identifying and helping make decisions on borderline cases, where probabilities are unclear, and based on the latent space projections they can find similar reference cases. Finally, users suggested that the auto-generated reports are useful to be attached to the patients' electronic health records.

Regarding the system's ease of use, users generally expressed their satisfaction, considering the user interaction fluid and comprehensible. They suggested that integrating the CDSS in the picture archiving and communication system (PACS) of the hospital would significantly increase its efficiency during routine practice.

Discussion

In this study, the use of the CDSS as an additional radiological information procedure provided an apparent better performance. However, for this sample size it was not significant, according to the resultant confidence intervals and statistical tests of differences. Besides, the positive difference in REC_M metric is suggestive of the system's reliability in imbalanced problems. As a pilot study, we must consider these outcomes as suggestive of the improvement in diagnostic accuracy on real settings. A larger sample size would be required to confirm.

The results suggest that when Curiam BT is used, an apparent larger improvement effect in terms of REC_M is observed, in comparison with ACC. It may indicate that using conventional MR methods the participant novice radiologists tend to diagnose cases as the more prevalent classes; however, when using the CDSS they are more confident in making diagnoses of less prevalent classes. Probably as a consequence of this, users trusted the system outcome as a diagnostic confirmation tool and to make decisions in borderline cases.

We compare next our work with previously published evaluations involving CDSS for brain tumour diagnosis ([4], [9] and [10]). As a main difference with respect to their approaches, we have focused our work to study the impact on novice radiologists. Also, following the Fuster et al. study [7], that

demonstrates the feasibility of classifying 3.0*T* cases with 1.5*T*-based predictive models, we have been the first to include both 1.5 and 3.0*T* cases in the evaluation.

In [4], [9] and [10], they evaluated 16, 10 and 40 patient cases respectively. However, they distributed the set of cases among the different radiologists, which increased the total number of evaluated cases, although involving repetition. Also, they counted with a larger number of radiologists to evaluate cases. In our study we randomly separated our 55 patient cases into the experiment and control groups. The limited resources we counted with did not permit distributing all the cases among all the evaluators, which may have provided other interesting measures such as evaluating the effect on inter-evaluator agreement. We may also have applied other solutions such as a prior stratification of cases based on a previous reference diagnosis. However, we can observe that in general our results are in line with previous works, where the use of MRS-based CDSS may improve the radiologic diagnostic efficacy.

We can discuss other differences in the methods. The other works evaluate classical area under the receiver operating characteristic (ROC) curve –based measurements on multiple-class settings. It requires obtaining SEN and SPE for each class, which entails grouping the other classes into a negative class (dichotomization). Fawcett [26] denominates this as class-reference formulation. He states that although this approach may alter ROC measurements on multiple-class settings it still provides a valid measurement. Solokova and Lapalme, however, consider that <<there is yet no well-developed multi-class ROC analysis>>. To our view, we consider that this approach may be problematic when the number of cases is low. It is mainly due to 1) the possible change in class prevalence and 2) the possibility that the dichotomized negative class contain erroneous predictions among its internal real classes (e.g., if we dichotomize class 1 vs. [2,3], misclassifications of 2 as 3 or 3 as 2 will be counted as TN_1). For this reason, we consider that ACC and REC_M (and also REC or PREC when the number of cases is large) are the most suitable measurements of quality for these problems, since they do not require dichotomizing: as shown in equations 1 and 2, each summand measures the number (ACC) or proportion (REC_M) of correct answers of those cases labelled as *c*, not needing a complementary group. In addition, a better measurement would be a cost-based evaluation, where each correct or incorrect classification involves different benefits or costs (i.e., a cost matrix). However, it is not straightforward to define those costs in clinical domains.

In the previous section we have listed the set of new predictive models suggested by users to be included in the system, with the purpose of improving its clinical usefulness. Given the difficulty to solve some of these classifications with MRS, we suggest studying multiparametric approaches from MR techniques (i.e., diffusion and perfusion images) following the quantitative MR methodology [14].

We also mentioned that users suggested integrating the system in the hospital PACS. The main reason of this can be that as an external tool to the PACS, it was required to gather the RAW MRS files from the scanner to load them in the CDSS, entailing an additional time in the workflow that, regardless of the possible efficacy benefits, may reduce the overall efficiency.

Results of the qualitative evaluation show that users were satisfied with the usability of Curiam BT. The experience in the development of CDSS by the authors and, specially, the user requirements defined based on the knowledge acquired during the European Projects HealthAgents and eTumour may have contributed to that. However, these requirements were defined by expert radiologists. Thus, as evaluated by novice radiologists in this study, we found that this other type of users may require an adaptation to the use of MRS, slightly varying the requirements, what was reflected in the suggestion of facilitate an automatic MRS processing method.

Finally, the initial results in improvement of diagnostic performance as well as the estimated number of samples for obtaining significant differences, suggest developing a larger RCT focusing to the discrimination of tumour grades as well as low prevalent classes. Additionally, the provided feedback regarding other clinically relevant questions is in line with this proposal, establishing the basis for a large clinical trial.

Conclusions

Despite the improved diagnostic accuracy and REC_M achieved by using Curiam BT, this study should be interpreted as a pilot experiment to assess the feasibility of incorporating the CDSS in a

radiological service and to plan a large clinical trial. The results of this work**,** with the novelty of evaluating both 1.5 and 3.0*T* cases, are in line with evaluations of similar systems previously published. In addition, we have compiled a large feedback to improve the clinical usefulness and quality of CDSSs for brain tumour diagnosis, which can give insights to researchers in further works. Finally, in this pilot study we have demonstrated that Curiam BT can be a useful tool for the training of novice radiologists in the diagnosis of brain tumours. In further work, the compiled feedback will be used to prepare an improved version of Curiam BT for its integration in the clinical workflow.

Acknowledgements

This work has been supported by the Spanish Ministry of Science and Innovation - Instituto de Salud Carlos III - FIS contract PI09/90177; and Universitat Politècnica de València – INNOVA UPV 2008 2 1834. We specially thank Miguel Ángel Edo, María Vañó, Carmen Barber, Ana Català-Gregori, Enrique Mollá and Cecilio Poyatos, for their collaboration in the development of this study.

References

[1] The eTUMOUR Consortium. eTumour, Web accessible MR decision support system for brain tumour diagnosis and prognosis, incorporating in vivo and ex vivo genomic and metabolomic data. VI Framework Programme, EC, Tech Rep FP6-2002-LSH-503094.

[2] Howe FA, Opstad KS. 1H MR spectroscopy tumours and masses. NMR Biomed 16(3) (2003) 123–131

[3] Galanaud D, Nicoli F, Chinot O, Confort-Gouny S, Figarella- Branger D, Roche P, Fuentes S, Le Fur Y, Ranjeva JP, Cozzone PJ. Noninvasive diagnostic assessment of brain tumors using combined in vivo MR imaging and spectroscopy. Magn Reson Med 55(6) (2006) 1236–1245.

[4] Tate AR, Underwood J, Acosta D, Julià-Sapé M, Majós C, Moreno-Torres À, Howe FA, van der Graaf M, Lefournier V, Murphy MM, Loosemore A, Ladroue C, Wesseling P, Bosson JL, Cabañas ME, Simonetti AW, Gajewicz W, Calvar J, Capdevila A, Wilkins PR, Bell BA, Rémy C, Heerschap A, Watson D, Griffits JR and Arús C. Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. NMR Biomed. 19 (2006) 411-434.

[5] García-Gómez JM, Tortajada S, Vidal C, Julià-Sapé M, Luts J, Moreno-Torres À, Huffel SV, Arús C and Robles M. The effect of combining two echo times in automatic brain tumor classification by MRS. NMR in Biomedicine, 21 (10) (2008) 1112-1125.

[6] García-Gómez JM, Luts J, Julià-Sapé M, Krooshof P, Tortajada S, Robledo JV, Melssen W, Fuster-García E, Olier I, Postma G, Monleón D, Moreno-Torres A, Pujol J, Candiota AP, Martínez-Bisbal MC, Suykens J, Buydens L, Celda B, Van Huffel S, Arús C and Robles M. Multiproject-multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy. MAGMA 22 (1) (2009) 5-18.

[7] Fuster-Garcia W, Navarro C, Vicente J, Tortajada S, García-Gómez JM, Sáez C, Calvar J, Griffiths J, Julià- Sapé M, Howe F, Pujol J, Peet AC, Heerschap A, Moreno-Torres A, Martínez-Bisbal MC, Martínez-Granados B, Wesseling P, Semmler W, Capellades J, Majós C, Alberich-Bayarri A, Capdevila A, Monleón D, Martí-Bonmatí L, Arús C, Celda B, and 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. Magnetic Resonance Materials in Physics, Biology and Medicine, 24(1) (2011) 35–42.

[8] Sáez C, García-Gómez JM, Vicente J, Tortajada S, Fuster E, Esparza M, Navarro A and Robles M. Curiam BT, Decision Support System for Brain Tumour Diagnosis. European Society for Magnetic Resonance in Medicine and Biology Congress (ESMRMB 2009), EPOS Posters/Paper Posters/Info-RESO:538 (2009).

[9] Celda B, Cano JG, Martinez-Bisbal M, Martinez-Granados B, and eTUMOUR-partners. Clinical evaluation of a fully automated Computer Aid Decision System (CADS) for brain tumour supported diagnosis. ISMRM 2009 Congress (2009).

[10] Julià-Sapé M, Coronel I, Majós C, Candiota AP, Serrallonga M, Cos M, Aguilera C, Acebes JJ, Griffiths J and Arús C. Prospective diagnostic performance evaluation of single-voxel 1H MRS for typing and grading of brain tumours. NMR Biomed. 25 (2012) 661-673.

[11] Magnetic Resonance User Interface, http://www.mrui.uab.es/mrui/ (Last accessed: 2013/07/22).

[12] Sáez C, García-Gómez JM, Vicente J, Tortajada S, Luts J, Dupplaw D, Van Huffel S and Robles M. A generic and extensible automatic classification framework applied to brain tumour diagnosis in HealthAgents. The Knowledge Engineering Review; 26(Special Issue 03) (2011) 283–301.

[13] Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, Scheithauer BW, Kleihues P. The 2007 WHO Classifcation of Tumours of the Central Nervous System. Acta Neuropathol 114 (2007) 97–109.

[14] Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 13(3) (1989) 319–339.

[15] Eta S. Berner (Editor). Clinical Decision Support Systems: Theory and Practice. Springer. ISBN: 978-0387339146 (2006).

[16] Hunt DL, Haynes R, Hanna SE and Smith K. Effects of Computer-Based Clinical Decision Support Systems on Physician Performance and Patient Outcomes: A Systematic Review. JAMA. 280(15) (1998) 1339-1346.

[17] Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, Sam J and Haynes RB. Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review. JAMA 293(10) (2005) 1223-1238.

[18] González-Vélez H, Mier M, Julià-Sapé M, Arvanitis T, García-Gómez JM, Robles M, Lewis P, Dasmahapatra S, Dupplaw D, Peet A, Arús C, Celda B, Huffel SV, Lluch i Ariet M. HealthAgents: distributed multi-agent brain tumor diagnosis and prognosis. Appl. Intell. 30(3) (2009) 191-202.

[19] Horská A and Barker PB. Imaging of brain tumors: MR spectroscopy and metabolic imaging. Neuroimaging Clin N Am. 20(3) (August) (2010) 293-310.

[20] Sáez C, García-Gómez JM, Vicente J, Tortajada S, Esparza M, Navarro AT, Fuster E, and Robles M. A generic decision support system featuring an assembled view of predictive models for magnetic resonance and clinical data. European Society

for Magnetic Resonance in Medicine and Biology Congress (ESMRMB 2008), Valencia, EPOS Posters/Paper Posters/Info-RESO:483 (2008).
[21] Vicente J, Sáez C, Tortajada S, Fuster-Garcia E, Esparza M, Robles M, and García-Gómez JM. Curiam BT kids, a Clinical DSS for pediatric brain tumour diagnosis. European Society for Magnetic Resonance in Medicine and Biology, Lisbon, Portugal, 25:622 (October) (2012).
[22] Tortajada S, García-Gómez JM, Vicente J, Sanjuán J, de Frutos R, Martín-Santos R, García-Esteve L, Gornemann I, Gutiérrez-Zotes A, Canellas F, Carracedo Á, Gratacos M, Guillamat R, Baca-García E, and Robles M. Prediction of postpartum depression using multilayer perceptrons and pruning. Methods of Information in Medicine, 48(3) (2009) 291–298.
[23] Agresti A and Coull B.A. Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions The American Statistician, 52 (2) (1998) 119-126.
[24] Sáez C, García-Gómez JM, Alberich-Bayarri Á, Edo MA, Vañó M, Català-Gregori A, Barber C, Poyatos C, Mollá E, Martí-Bonmatí L, Robles M. Clinical validation of the added value of a clinical decision support system for brain tumour diagnosis based on SV 1H MRS: randomized controlled trial of effectiveness and qualitative evaluation. Short oral communication presented at the 24th International Conference of the European Federation for Medical Informatics. Pisa, Italy (August) (2012).
[25] Wei Q, Dunbrack RL Jr The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. PLoS ONE 8(7): e67863. (2013).
[26] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters 27 (2006) 861–874.
[27] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45 (4) (2009) 427-437.