Universitat Autònoma de Barcelona

HOSPITAL DE LA SANTA CREU I SANT PAU
UNIVERSITAT AUTÒNOMA DE BARCELONA

Programa de Doctorat en Medicina (RD 1393/2007)

Departament de Medicina

Automatic image quantification strategies in clinical nuclear medicine and neuroradiology

Tesis Doctoral presentada per

Frederic Sampedro Santaló

per a obtenir el grau de Doctor per la
Universitat Autònoma de Barcelona

Sota la direcció de:

Prof. Ignasi Carrió Gasset

Dr. Sergio Escalera Guerrero

Dr. Jordi Riba Serrano

Abstract

With the revolution of digital medical imaging and the increasing computational power, the field of quantitative medical image analysis emerged. By programming a computer to detect patterns of interest in medical images and derive clinically meaningful numerical indicators from them, this field shows promising potential for healthcare and medical research systems.

In this thesis, the design and implementation of computer-based quantification techniques in nuclear medicine and neuroradiological images led to several contributions in this field. These image-derived indicators contributed to complement the visual diagnosis and to further understand the pathophysiology of important health issues such as breast cancer, non-Hodgkin lymphoma, pyelonephritis, Alzheimer's disease, Parkinson's disease and cannabis abuse.

Resum

Amb la revolució de la tecnologia digital d'obtenció d'imatges radiològiques i l'increment de la potència computacional, el camp de la quantificació d'imatges mèdiques ha sorgit. El fet de poder programar un ordinador per a que detecti patrons d'interès en imatges radiològiques i pugui derivar-ne d'aquests indicadors numèrics amb valor clínic fa que, sens dubte, aquest àmbit de coneixement tingui un gran potencial en entorns mèdics i de recerca.

En aquesta tesi es presenten un conjunt de contribucions científiques en aquest context. En particular, es descriu el disseny i la implementació d'una sèrie d'estratègies computacionals de quantificació d'imatges de medicina nuclear i neuroradiologia. A continuació es detalla com aquestes tècniques han demostrat ser d'utilitat per a l'estudi de malalties molt rellevants en l'actualitat com són el càncer de mama, el limfoma no-Hodgkin, la pielonefritis, la malaltia d'Alzheimer, la malaltia de Parkinson i l'abús de cànnabis.

Resumen

Con la revolución de la tecnología digital de obtención de imágenes radiológicas y el aumento de la potencia computacional, el campo de la cuantificación de imágenes médicas ha emergido. El hecho de poder programar un ordenador para que detecte patrones de interés en imágenes radiológicas y pueda derivar de ellos una serie de indicadores numéricos con valor clínico hace que, sin duda, este ámbito de conocimiento tenga un gran potencial en el entorno médico y de investigación.

En esta tesis se presentan un conjunto de contribuciones científicas en este contexto. En particular, se describe el diseño y la implementación de una serie de estrategias computacionales de cuantificación de imágenes de medicina nuclear y neuroradiología. A continuación se detalla cómo estas técnicas han demostrado ser de utilidad en el estudio de patologias muy relevantes en la actualidad como son el cáncer de mama, el linfoma no-Hodgkin, la pielonefritis, la enfermedad de Alzheimer, la enfermedad de Parkinson i el abuso de cánnabis.

A la meva mare i,

especialment,

al meu pare.

"In theory, there is no difference between theory and practice. But, in practice, there is."

Attributed to multiple people. It's so true that it doesn't matter who said it.

Acknowledgements / Agraïments / Agradecimientos

En primer lloc agrair el suport incondicional dels meus pares, que ha estat un pilar fonamental per al desenvolupament d'aquesta tesi.

En segon lloc, agrair al Prof. Ignasi Carrió el fet de donar-me l'oportunitat de realitzar aquesta tesi sota la seva supervisió i, a més, permetre'm "volar sol". Gràcies a ell he pogut organitzar-me el temps amb total llibertat, he pogut compaginar els estudis amb la tesi, i he tingut l'oportunitat de conèixer i col·laborar amb nombrosos grups de recerca de l'Hospital de Sant Pau.

En tercer lloc, agrair al Dr. Sergio Escalera el seu esforç i ajuda en els primers passos d'aquesta tesi, que van ser absolutament claus per endinsar-me amb bon peu en el món de la recerca.

Finalment, el meu sincer agraïment a totes aquelles persones que han col·laborat en fer possible aquesta tesi, especialment als membres del servei de Medicina Nuclear (Dra Anna Domenech, Dra Valle Camacho), de la Unitat de Memòria (Dr Juan Fortea, Eduard Vilaplana), de la Unitat de Trastorns del Moviment (Saül Martínez-Horta, Dr Jaume Kulisevsky), del grup de Neuropsicofarmacologia Humana de l'Hospital de Sant Pau (Dr Jordi Riba) i del departament de matemàtica aplicada i análisis de la UB (Dra Anna Puig).

Index

I. List of publications

1. *Automatic Tumor Volume Segmentation in Whole-Body PET/CT Scans: A Supervised Learning Approach. F. Sampedro, S. Escalera, A. Domenech,I. Carrio. J. Med. Imaging Health Inf. 5, 192-201,2015.*

2. *Obtaining quantitative global tumoral state indicators based on whole-body PET/CT scans: a breast cancer case study. F.Sampedro, A. Domenech, S.Escalera. Nuclear Medicine Communications 35(4),362-371, 2014.*

3. *A computational framework for cancer response assessment based on oncological PET-CT scans. F. Sampedro, S.Escalera, A.Domenech, I.Carrio. Computers in Biology and Medicine 55, 92-99, 2014.*

4. *Deriving global quantitative tumor response parameters from 18F-FDG PET-CT scans in patients with non-Hodgkin's lymphoma. F.Sampedro, A.Domenech, S.Escalera, I.Carrio. Nuclear Medicine Communications 36 (4), 328-333, 2015.*

5. *Computing quantitative indicators of structural renal damage in pediatric DMSA scans. F. Sampedro, A.Domenech, S.Escalera, I.Carrió. Rev Esp Med Nucl Imagen Mol, In Press, 2016.*

6. *APOE-by-sex interactions on brain structure and metabolism in healthy elderly controls. Sampedro F, Vilaplana E, de Leon MJ, Alcolea D, Pegueroles J, Montal V, Carmona-Iragui M, Sala I, Sánchez-Saudinos MB, Antón-Aguirre S, Morenas-Rodríguez E, Camacho V, Falcón C, Pavía J, Ros D, Clarimón J, Blesa R, Lleó A, Fortea J; Alzheimer's Disease Neuroimaging Initiative. Oncotarget. 05 Sep 9;6(9):666-74,2015.*

7. *Non-demented Parkinson's disease patients with apathy show decreased grey matter volume in key executive and reward-related nodes. S.Martínez-Horta & F.Sampedro, J. Pagonabarraga, J. Marín-Lahoz, J.Riba, J.Kulisevsky. Brain Imaging and Behavior, In Press, 2016.*

8. *Telling true from false: Cannabis users show increased susceptibility to false memories.J. Riba, M. Valle & F. Sampedro, A. Rodríguez-Pujadas, S. Martinez-Horta, J. Kulisevsky, A.Rodriguez-Fornells. Molecular Psychiatry 20, 772–777,2015.*

II. Thesis introduction, motivation and background

Digital medical imaging has been one of the revolutions in medicine of the last two decades. Medical specialists currently have available a wide range of imaging modalities that have raised the diagnostic quality and follow-up evaluation of many pathologies.

Most medical images used in clinical practice are visually evaluated by the trained physician in order to detect and characterize the presence of a particular pathology of interest. While in most cases this procedure is assumed to obtain the best accuracy, it has several imitations. First, the diagnostic performance is highly dependent on the physician's expertise. Second, the diagnostic product is generally categorical (i.e. positive/negative/inconclusive) and descriptive, lacking a quantitative modeling of the characteristics underlying a disorder. Finally, some image modalities cannot be directly evaluated visually due to their technical nature (e.g. resting-state cerebral functional magnetic resonance imaging), requiring image-derived quantitative indicators to be properly assessed.

To overcome these limitations and given the increasing computational power of modern technology, the automatic computation of quantitative indicators in medical images that could complement the visual evaluation by the trained physician is emerging as an important research area. Notably, the information from this type of indicators would be, by definition, observer-independent and quantitative.

Therefore, for each medical scenario and image modality of interest, the challenge of designing the best computational image-quantification strategy capable of obtaining new indicators that could improve diagnostic accuracy, prognosis estimation or disease understanding is clearly appealing to the medical community.

The possible incorporation of this type of indicators in healthcare centers would have major advantages at several levels. At the clinical level, if the image-derived quantitative indicators aided in the interpretation of complex radiological patterns by providing relevant observer-independent diagnostic information, they would contribute to a better overall diagnostic accuracy, especially in situations where there is limited physician expertise on this task. At the management level, if the availability of such indicators contributed to accelerate the determination of radiological conclusions, an increase in diagnostic throughput would be obtained, thus providing significant management and economic benefits to the health institution.

III. Objectives

The main objective of this thesis is to provide scientific contributions to the field of automatic medical image quantification for clinical or research applications. Among the large number of medical imaging modalities and clinical contexts, this thesis will only focus on the computation of quantitative indicators from the following image modalities: FDG-PET scans (whole-body and cerebral), renal DMSA-scans, T1-weighted magnetic resonance imaging (cerebral), and cerebral event-related functional magnetic resonance imaging. Details about these image modalities are given throughout this section.

The indicators computed from these images were designed to contribute to application-specific scenarios in several medical domains, including tumor burden evaluation, structural kidney damage quantification, neurodegenerative disease characterization and drug-induced brain activation analysis.

In this section, each of these image quantification contexts is described and the original research papers derived from them are mentioned, whereas the next section will summarize its main results and discussions.

The first automatic image quantification scenario of this thesis was related to tumor burden characterization in oncological whole-body FDG-PET scans. This type of scan is a valuable nuclear medicine test where a whole-body image of the subject is obtained, showing the uptake distribution of a glucose-analog radioactive tracer. The diagnostic power of this imaging technique relies on the fact that elevated tracer uptake (a proxy for high metabolic activity) in certain anatomical locations may indicate the presence of a tumor.

In this first image analysis context, the main objective was to obtain a numerical indicator from FDG-PET scans that was able to model underlying tumor properties (i.e. volume, metabolic uptake and spread) in a quantitative and expert-independent manner (Fig.1).

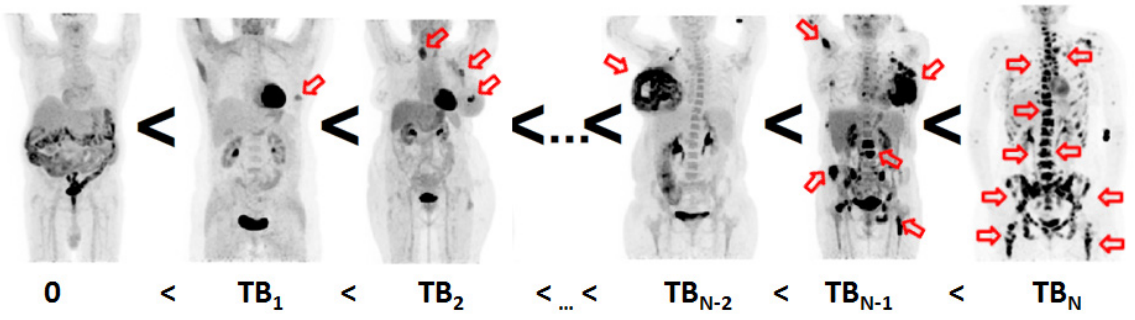


Fig.1 First scenario: automatic computation of an efficient image-derived tumor burden indicator (TBI) in whole-body FDG-PET scans to model quantitatively the spectrum of tumor states in the population. Red arrows indicate the presence of tumor volume in that anatomical region.

From the technical point of view, a key step needed to automatically compute this type of indicator is that a computer algorithm should be able to detect and segment the tumor

volume of *any* given whole-body FDG-PET scan. Given the substantial variability of the human anatomy and physiological tracer uptake of the population in this context, this poses *per se* a major computational problem.

Therefore, the first contribution (article [1]) sought to apply state-of-the-art machine learning and computer vision algorithms to automatically detect and segment tumor volume in this type of scans. Then, having the tumor volume segmented from the images either by using automatic or expert-guided segmentation techniques, the scenario of computing accurate numerical indicators that would be able to model the underlying tumor characteristics was addressed in article [2]. Both of these works used a cohort of breast cancer patients where novel quantitative indicators could contribute to obtain image-derived prognostic indicators of the disease.

In the same image modality context but addressing a different clinical scenario, the computation of quantitative indicators modeling the tumor response in time from a pair of pre- and post- treatment FDG-PET scans was addressed; both at the technical and clinical levels (articles [3] and [4]). In this case, a cohort of non-Hodgkin lymphoma patients with a pair of time-consecutive FDG-PET scans was used to derive new indicators to quantify the tumor progression or response.

The second image quantification context sought to detect and characterize structural renal damage in DMSA scans. This type of scan is also a nuclear medicine image where the radiotracer used shows the distribution of the DMSA molecule within the kidneys. This distribution is of particular diagnostic interest since healthy kidney tissue would tend to show high tracer uptake, whereas damaged kidney regions would show less or no tracer uptake.

The automatic computation of numerical indicators in this type of image (Fig.2) led to the contribution presented in article [5], where a pediatric cohort of patients with structural kidney damage was used to obtain image-derived quantitative indicators of the underlying renal pathology.

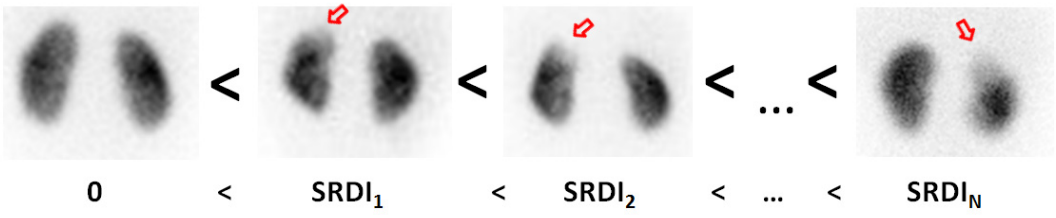


Fig.2 Second scenario: automatic computation of an image-derived efficient structural renal damage indicator (SRDI) in DMSA scans to model quantitatively the underlying pathology. Red arrows indicate structural renal damage in those kidney areas.

The third image analysis scenario aimed to quantify the metabolic activity from cerebral FDG-PET scans (described above) in a set of key brain areas related to dementia such as the temporal lobe (Fig.3). In this case, the relationship between such imaging indicators, the APOE genotype (the best-known genetic risk factor for Alzheimer's disease) and gender in healthy controls was investigated in article [6].

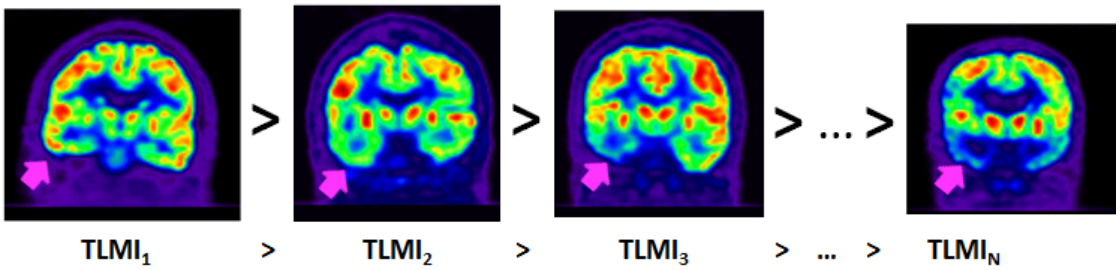


Fig 3 Third scenario: automatic computation of an efficient temporal lobe metabolism indicator (TLMI) in cerebral FDG-PET scans to model quantitatively the underlying pathophysiology of Alzheimer's disease. Pink arrows illustrate the metabolic activity of a region in the left temporal lobe.

The fourth and fifth image quantification contexts were performed on cerebral T1-weighted magnetic resonance maging (MRI) scans. By applying a combination of magnetic fields and radio frequency pulses to the brain, tissues with different properties generate different signals that can be detected and processed, obtaining a structural brain image where gray matter, white matter, and cerebrospinal fluid can be distinguished.

The amount and distribution of gray matter in the brain is key to correct brain functioning. An accurate measure of the integrity of gray matter within the cerebral cortex (responsible for cognition and memory among other important functions) is cortical thickness, defined by the distance between the white matter and pial surfaces. For subcortical brain structures such as the basal ganglia, gray matter volume can be computed by direct segmentation of the volumetric image.

Cortical atrophy (i.e. reduction of cortical thickness) is one of the hallmarks of Alzheimer's disease, whereas alterations in the gray matter volume of the basal ganglia are associated with movement disorders (Fig. 4).

The work described in article [6] also addressed the quantification of cortical thickness in key areas of the cerebral cortex to also analyze the influence of APOE and gender on brain atrophy associated with the Alzheimer's disease.

Quantification of gray matter volume in a specific region of the basal ganglia known as the nucleus accumbens (a key node of the brain's reward circuit) was performed in Parkinson's disease patients with apathy. This aimed to provide a neuroanatomical basis for this behavioral manifestation, which plays an important role as a marker of disease progression (article [7]).

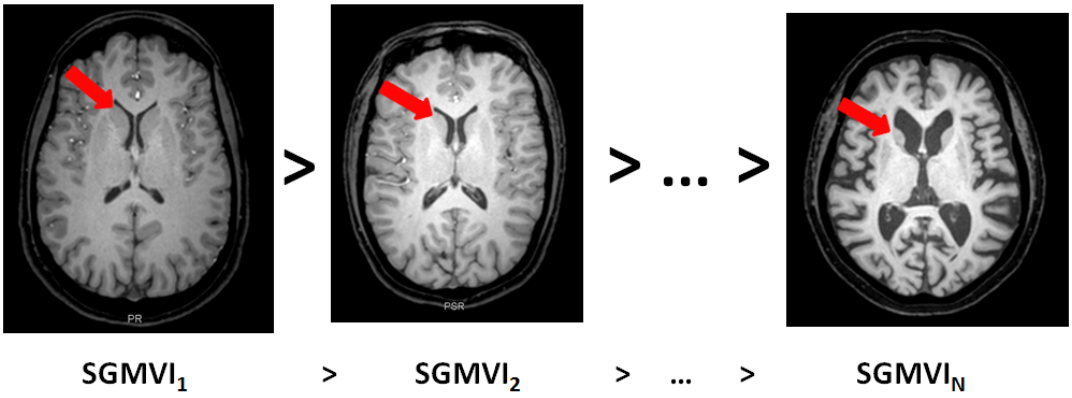


Fig. 4. Fifth scenario: automatic computation of a striatal gray matter volume indicator (SGMVI) in T1-MRI images to model quantitatively the underlying pathophysiology of Parkinson's disease. Red arrows illustrate the gray matter volume of a portion of the left caudate nucleus.

The sixth and last image analysis scenario was related to the quantification of functional magnetic resonance (fMRI) images. Neuronal activation requires energy, and therefore the vascular system provides nutrients and oxygen to those neurons that have increased activation. This regional change of the vascular content and flow in the regions where there is a high neuronal activity can be detected using Blood-oxygen-level dependent (BOLD) MRI signal.

By acquiring a set of BOLD fMRI images for a period of time, the brain activity during specific events can be recorded. If a particular task is performed during the acquisition, the activation of a particular brain region of interest related to that specific task can be computed (Fig.5).

In article [8], cortical and hippocampal activation during a false memory rejection task was quantified from the BOLD images in a set of cannabis users and healthy controls. Areas of differential activation between groups were identified. Activity in these brain regions were correlated with individual values of lifetime cannabis use. Results identified novel memory impairments in users and increased our understanding of the deleterious impact of cannabis on cognition.

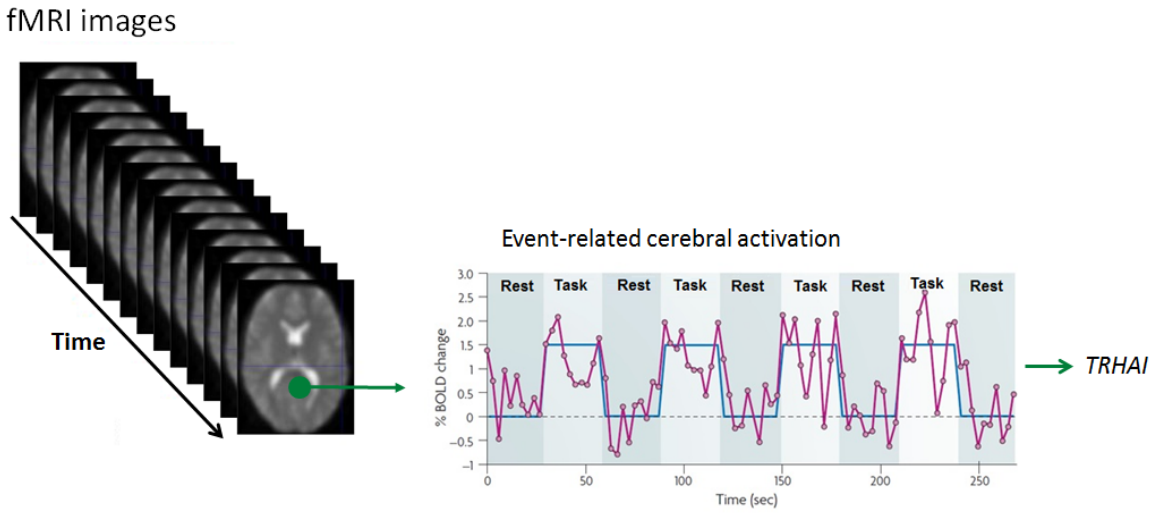


Fig. 5. Sixth scenario: automatic computation of a task-related hippocampal activation indicator (TRHAI) from event-related fMRI images to quantitatively model the possible memory alterations in cannabis users.

IV. Summary of the contribution's results and discussion

In this section, a brief summary of the main results and discussion of each of the works introduced in the previous section is presented.

Automatic Tumor Volume Segmentation in Whole-Body PET/CT Scans: A Supervised Learning Approach. F. Sampedro, S. Escalera, A. Domenech,I. Carrio. J. Med. Imaging Health Inf. 5, 192-201,2015.

In this work the application-specific problem of automatic detection and segmentation of tumor volume in whole-body FDG-PET/CT scans was addressed.

The visual inspection by the medical specialist of whole-body FDG-PET/CT scans is a valuable diagnostic tool in oncological scenarios. Furthermore, the possibility of quantifying a set of tumor properties (such as its total volume or metabolic uptake) using expert-guided segmentation tools offers additional clinical value. However, this expert-guided segmentation process is highly time-consuming and expert-dependent. This work addressed the need to automate this task and proposed a computational system to do so.

By applying state-of-the-art machine learning techniques (Multiscale-Stacked Sequential Learning), an automatic FDG-PET tumor segmentation system was obtained. Machine learning techniques are a type of artificial intelligence algorithms that learn how to accomplish a particular task if they are provided with a set of training examples. In this case, expert-guided tumor segmentations of 100 breast cancer FDG-PET scans were used to let the system learn a set of segmentation rules.

The segmentation results of the automatic system achieved, at the pixel/voxel level, on average 49% sensitivity, 99% specificity and 39% Jaccard Overlap Index (a measure of comparison of automatic vs. expert-derived segmentation results). Furthermore, the total tumor volume of the breast cancer scans was computed from both expert-guided and automatic segmentation outputs, showing a correlation of 73%.

Conceptually, these results show that the obtained automatic segmentation system successfully detected and segmented only very clear and reasonably-sized tumor regions within the scans. They also suggest that even though the automatic segmentation accuracy at the pixel-level may not be equivalent to that obtained by the manual segmentation performed by a trained physician, at the global indicator level (such as the whole body total tumor volume), the automatically-computed indicators would have potential diagnostic value within the clinical environment.

Obtaining quantitative global tumoral state indicators based on whole-body PET/CT scans: a breast cancer case study. F.Sampedro, A. Domenech, S.Escalera. Nuclear Medicine Communications 35(4),362-371, 2014.

In this work, the need to find a set of quantitative indicators derived from the image analysis of whole body FDG-PET/CT scans aiming to model properly the global oncological state of the patient was addressed.

A set of 100 breast cancer FDG-PET/CT scans were classified according to their global oncological state following visual evaluation by a consensus of nuclear medicine physicians. A set of quantitative indicators derived from the tumor segmentation of the images was then computed, both using completely automatic and expert-guided approaches. The performance of these indicators at modeling the patient's underlying oncological state as classified by the expert's visual inspection was measured.

The performance results of the commonly used indicators in clinical practice including whole-body metabolic tumor volume (WBMTV), maximum/mean metabolic activity of the tumor (SUVmax/SUVmean), and a combination of both (Total Lesion Glycolysis) achieved performances ranging from 49% to 79% (in a measure of correlation with the expert's classification).

None of these indicators take into account the spread of the tumor across the patient's body (i.e. the number of anatomical structures where the tumor is present). This work proposed to include that information into the computation of the indicators, obtaining new indicators that improved performance from 80 to 87%. These results were obtained using expert-guided tumor segmentations. Using a completely automatic approach, the best performance result was 64%.

Taken together, this work contributed to show that image-derived FDG-PET global quantitative indicators can prove useful in clinical nuclear medicine and oncological scenarios. Furthermore, the incorporation of tumor spread measures in the computation of such indicators improved its performance at modeling the underlying oncological state. Finally, a substantial performance difference between the expert-guided and the completely automatic computation of the indicators was observed, suggesting that there is room for improvement in this research line.

A computational framework for cancer response assessment based on oncological PET-CT scans. F. Sampedro, S.Escalera, A.Domenech, I.Carrio. Computers in Biology and Medicine 55, 92-99, 2014.

In the following two works, a different clinical scenario within the whole-body PET-CT image analysis context was addressed: the quantification of oncological state changes over time.

This first contribution focused exclusively on the design and implementation of a computational framework aimed to correctly identify the most common clinical cancer evolution scenarios (i.e. progression, partial response, total response, mixed response, and relapse) from a pair of time consecutive PET-CT scans of the patient. This task, visually performed by nuclear medicine physicians in clinical practice (suffering from the common expert-dependence and qualitative diagnostic product limitations), poses a challenging problem within a computational environment.

Performance results at predicting the cancer evolution scenario in a set of 100 non-Hodgkin lymphoma (NHL) patients achieved up to 90% of accuracy when using expert-guided image-derived tumor segmentations and 70% accuracy when using a completely automatic approach. These results suggest that computing a set of image-derived quantitative indicators of cancer dynamics is only reasonable if expert-derived tumor segmentation information is available.

Deriving global quantitative tumor response parameters from 18F-FDG PET-CT scans in patients with non-Hodgkin's lymphoma. F.Sampedro, A.Domenech, S.Escalera, I.Carrio. Nuclear Medicine Communications 36 (4), 328-333, 2015.

Based on the results of the previous study, this work focused on the actual computation of image-derived quantitative indicators designed to model the magnitude of cancer response or progression conditions.

A set of 89 pairs of time consecutive PET-CT scans presenting NHL were classified by a consensus of nuclear medicine physicians into progressions, partial responses, mixed responses, complete responses, and relapses. The cases of each group were ordered by magnitude following visual analysis. Thereafter, a set of quantitative indicators designed to model the cancer evolution magnitude within each group were computed using expert-guided and automatic image-processing techniques. Performance evaluation of the proposed indicators was measured by a correlation analysis with the expert-based visual analysis.

The set of proposed indicators achieved the following correlation results in each group with respect to the expert-based visual analysis: 80.2% in progressions, 77.1% in partial responses, 68.3% in mixed responses, 88.5% in complete responses, and 100% in relapses. In the progression and mixed response groups, the proposed indicators outperformed the common indicators used in clinical practice (i.e. changes in WBMTV, SUVmax, SUVmean, and total lesion glycolysis) by more than 40%. These results were

obtained using expert-guided tumor segmentations. In this scenario, the automatic approach obtained very poor performance results (<30%).

These results show that the computation of global indicators of NHL response using PET-CT imaging techniques offers a strong correlation with the associated expert-based visual analysis, motivating the future incorporation of such quantitative and highly observer-independent indicators in oncological decision making or treatment response evaluation scenarios. However, a robust automatic approach to the computation of such indicators is still to be obtained.

Computing quantitative indicators of structural renal damage in pediatric DMSA scans. Rev Esp Med Nucl Imagen Mol, In Press, 2016.

The aim of this work was to propose, implement and validate a computational DMSA quantification framework for the computation of image-derived indicators that seek to model the underlying structural renal damage in a quantitative and observer-independent manner.

With this objective in mind, a set of image-derived quantitative indicators based on the relative lesion's size, intensity and histogram distribution was computed from a set of 16 pediatric DMSA-positive scans and 16 matched controls, using both expert-guided and automatic approaches. A correlation analysis was conducted to investigate the association of these indicators with other clinical data of interest in this scenario, including C-reactive protein (CRP), leukocyte count, vesicouretral reflux, fever, relative perfusion, and the presence of renal sequelae in a 6-month follow-up DMSA scan.

A fully automatic lesion detection and segmentation system successfully classified DMSA-positive scans from negative scans (AUC=0.92, sensitivity=81% and specificity =94%). The image-computed relative lesion size correlated with the presence of fever and CRP levels ($p<0.05$), and a measure derived from the histogram distribution of the lesion gave significant performance results in detecting permanent renal damage (AUC=0.86, sensitivity=100% and specificity=75%).

These results suggest that the proposal and implementation for the first time of a computational framework to quantify structural renal damage from DMSA scans shows promising potential to complement visual diagnosis and non-imaging indicators.

APOE-by-sex interactions on brain structure and metabolism in healthy elderly controls. Sampedro F, Vilaplana E, de Leon MJ, Alcolea D, Pegueroles J, Montal V, Carmona-Iragui M, Sala I, Sánchez-Saudinos MB, Antón-Aguirre S, Morenas-Rodríguez E, Camacho V, Falcón C, Pavía J, Ros D, Clarimón J, Blesa R, Lleó A, Fortea J; Alzheimer's Disease Neuroimaging Initiative. Oncotarget. 05 Sep 9;6(9):666-74.

In this work, the objective was to obtain image-derived quantitative indicators of hypometabolism and atrophy in some key areas of the brain related to Alzheimer's disease (AD), and relate them to a set of well-known risk factors of the disease.

In particular, the computation of glucose uptake and cortical thickness indicators within the brain's temporal lobe in a particular population of interest was addressed in order to understand a specific phenomenon observed at the clinical level related to the APOE4 genotype.

The APOE4 variant is the largest known genetic risk factor for late-onset sporadic AD. Epidemiologically, it has been shown that the APOE4 effect on Alzheimer Disease risk is stronger in women than in men. However, the underlying neural mechanisms of this observation had not been established. In this study, the APOE-by-sex interaction on brain metabolism, brain structure, and other indicators of interest such as cerebro-spinal fluid (CSF) was addressed.

This analysis was conducted in a sample of 328 healthy elderly controls from the Alzheimer's Disease NeuroImaging initiative database. Focusing on the brain metabolism and structure interaction results, sex stratification showed that female APOE4 carriers presented widespread brain hypometabolism and atrophy with respect to non-carriers. In contrast, APOE4 male carriers showed only a small region of hypometabolism and no atrophy with respect to non-carriers. This significant hypometabolic and atrophy pattern difference was especially prominent in the temporal lobe, a key brain region involved in AD ($p<0.001$).

These results suggest that the impact of APOE4 on brain metabolism and structure is strongly modified by sex, providing a biologically plausible explanation to the clinical observations. This finding should be taken into consideration in the interpretation of image-derived metabolic indicators commonly used in the clinical management of AD.

Non-demented Parkinson's disease patients with apathy show decreased grey matter volume in key executive and reward-related nodes. Brain Imaging and Behavior, In Press, 2016

In this work, the gray matter volume (GMV) quantification of a set of Parkinson's disease (PD) T1-MRI images contributed to understanding the neuroanatomical basis of apathy in this disease.

For this purpose, two groups of 18 PD patients with available T1-MRI scans were identified. Both groups were equivalent in terms of sociodemographic characteristics,

disease stage and treatment type. The groups only differed in the manifestation of apathy.

The quantification of GMV in cortical and subcortical structures showed that the apathetic group had reduced GMV in the nucleus accumbens ($p<0.005$), a key node of the brain's reward circuit, possibly explaining the motivational deficit observed in apathetic patients. Apathetic patients also had reduced GMV in regions involved in executive functions such as the orbitofrontal cortex ($p<0.005$). Moreover, the patients' GMV at those regions correlated with their cognitive performance ($p<0.001$).

These results suggest apathy as a marker of more widespread brain degeneration in Parkinson's disease.

Telling true from false: Cannabis users show increased susceptibility to false memories. J. Riba, M. Valle, F. Sampedro, A. Rodríguez-Pujadas, S. Martinez-Horta, J. Kulisevsky, A.Rodriguez-Fornells. Molecular Psychiatry (2015) 20, 772–777.

In this work the computation of event-related brain activation indicators was addressed. In particular, a group of 16 heavy cannabis users (which abstained from the drug at least 4 weeks prior to this study) and their matched controls performed a memory task within a functional magnetic resonance imaging (fMRI) scan.

The task performed within the fMRI station implemented a well-established method, the Deese-Roediger-McDermott paradigm, used to experimentally induce memory illusions or "false memories". In this task, "lure" stimuli have to be adequately identified and rejected by the participant.

Notably, cannabis users performed significantly poorer than the controls in this task: the number of incorrectly identified lure stimuli was higher in the cannabis-using group ($p<0.01$).

In order to obtain the neural correlates of this observation, quantitative indicators of brain activation during the task were computed in specific brain areas. Cannabis users showed widespread cortical hypoactivation following lure stimuli rejection. The brain areas involved were located in the frontal and parietal cortices and in the medial temporal lobe in a region including the hippocampus. These areas are known to be involved in executive control, attention and memory processes.

The fact that the hippocampal activation was significantly lower in the cannabis users ($p<0.001$) indicate a possible memory alteration. Importantly, the hippocampal activation in this group was inversely correlated with the lifetime amount of cannabis used by the subjects ($p<0.01$), further supporting the association of the brain's hypoactivation and the drug use.

These findings suggest that cannabis users have an increased susceptibility to memory distortions even when abstinent and drug-free, suggesting a long-lasting compromise of memory and cognitive control mechanisms involved in reality monitoring.

V. General conclusions and final remarks

This thesis presents a set of contributions to the field of automatic and quantitative analysis of digital medical images in nuclear medicine and neuroradiology. This expanding field aims to provide better diagnostic accuracy and help to detect subtle anatomical, physiological or pathological changes in clinical groups that may contribute to our understanding of disease etiologies.

On one hand, this work contributed to increasing the diagnostic potential of breast cancer and non-Hodgkin lymphoma FDG-PET scans through the design and computation of a set of novel image-derived quantitative indicators of tumor burden. In addition, quantitative analysis of pediatric DMSA scans helped to develop new indicators of structural renal damage.

On the other hand, the application-specific quantification of cerebral FDG-PET and T1-MRI scans contributed to the understanding of the pathogenesis of Alzheimer's and Parkinson's diseases, whereas the quantification of event-related fMRI images identified a novel memory deficit associated to cannabis use.

Taken together, these contributions illustrate the potential value of the design and implementation of ad-hoc automated quantification strategies of medical images. The availability of new imaging modalities sparks the search for optimal image-derived quantitative indicators that model each particular clinical context of interest. Such indicators can be used either to complement visual diagnosis or to aid in the comprehensive characterization of the underlying pathology.

VI. Original research papers

Journal of Medical Imaging and Health Informatics
Vol. 5, 192–201, 2015

Automatic Tumor Volume Segmentation in Whole-Body PET/CT Scans: A Supervised Learning Approach

Frederic Sampedro[1,*], Sergio Escalera[2], Anna Domenech[3], and Ignasi Carrio[3]

[1]*Autonomous University of Barcelona, Faculty of Medicine, 08193, Barcelona, Spain*
[2]*University of Barcelona, Faculty of Mathematics, Gran Via de les Corts 585, 08007, Barcelona, Spain*
[3]*Hospital de Sant Pau, Nuclear Medicine Department, 89 Carrer Sant Quintí, 08026 Barcelona, Spain*

Whole-body 3D PET/CT tumoral volume segmentation provides relevant diagnostic and prognostic information in clinical oncology and nuclear medicine. Carrying out this procedure manually by a medical expert is time consuming and suffers from inter- and intra-observer variabilities. In this paper, a completely automatic approach to this task is presented. First, the problem is stated and described both in clinical and technological terms. Then, a novel supervised learning segmentation framework is introduced. The segmentation by learning approach is defined within a Cascade of Adaboost classifiers and a 3D contextual proposal of Multiscale Stacked Sequential Learning. Segmentation accuracy results on 200 Breast Cancer whole body PET/CT volumes show mean 49% sensitivity, 99.993% specificity and 39% Jaccard overlap Index, which represent good performance results both at the clinical and technological level.

Keywords: PET/CT, Whole Body, Tumor Segmentation, Supervised Learning, Contextual Classification.

1. INTRODUCTION

^{18}F-FDG PET/CT (Fluorodeoxyglucose Positron Emission Tomography–Computed Tomography) is a nuclear medicine imaging technology widely used in cancer management. Its outputs are two registered 3D volumes: a PET scan, with low resolution, containing patient's metabolic information in SUV (Standard Uptake Value) units,[15] and a CT scan, with higher resolution, containing patient's anatomical information in Hounsfield (HU) units (Hoffer).

FDG-avid tumors show higher than normal SUV values in non-physiological locations. Given the low resolution of the PET scan, the CT scan is used to precisely locate any suspicious high metabolic activity within its anatomical context. Whole-body PET/CT scans are a valuable diagnostic and prognostic tool since they allow medical experts to evaluate the patient's global cancer stage, as well as to write a detailed descriptive report for each patient based on the observed findings.

Quantitative tumor information (Metabolic Tumor Volume, SUVmax, SUVmean, Total Lesion Glycolysis) obtained from PET scans has also proven to be useful in the clinical scenario, especially in follow-up scenarios.[21,29] In order to obtain these numerical parameters, "tumoral voxels" must be segmented from the PET volume (Fig. 1), identifying the patient's whole body metabolic tumor volume (WBMTV). In the clinical practice, this task is currently performed manually by trained physicians who, for each tumor lesion, run any convenient semi-automatic segmentation algorithm available (region growing, level-set or adaptive thresholding, among others) seeking for the correct region isolation.[26]

This process suffers from two main limitations. First, it is a time-consuming task, especially when working with advanced cancer stage patients, where the tumor has spread all over the whole body. Second, it suffers from inter-and intra-observer variabilities: aside from human related errors and variabilities, the fact of using different software tools, different semi-automatic segmentation algorithms or different initialization parameters for the same algorithm is likely to introduce non-desirable variations in the segmentation results.

In order to overcome these limitations, in this work we propose the implementation of a completely automatic whole-body PET tumor segmentation system. The proposed system will analyze the anatomical, physiological and physiopathological context of any given whole-body PET scan from breast cancer patients in order to be able to automatically identify, locate and segment the set of tumoral voxels present in it.

In order to deal with this task, a supervised learning framework is proposed, designed, implemented and validated on a 200 PET/CT ground truth dataset. The system is built using a

*Author to whom correspondence should be addressed.

 2156-7018/2015/5/192/010 doi:10.1166/jmihi.2015.1374

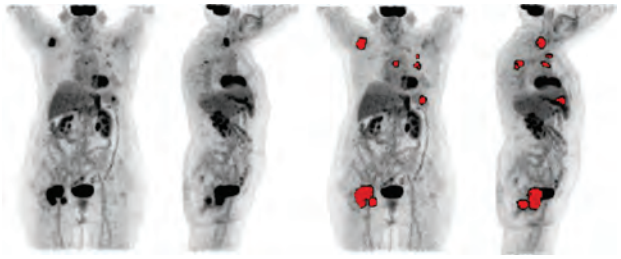


Fig. 1. Maximum intensity projections (MIP) of a sample PET scan (left) and its corresponding tumoral volume segmentation (right).

convenient set of voxel features, a cascade of AdaBoost classifiers and a 3D contextual information framework based on a modified proposal of Multi-scale stacked sequential learning strategy. The proposed system shows high performance results both at the clinical and technological level.

The rest of the paper is organized as follows. Section 2 describes the related work on this topic. Section 3 describes the dataset used in the design and validation of the proposed system. Section 4 fully describes the analysis, design and implementation of the proposed supervised learning system. Section 5 describes the system's performance results. Finally, Section 6 points out some conclusions and future work.

2. RELATED WORK

To the best of our knowledge, the particular problem of automatic whole body PET/CT metabolic tumor volume segmentation has not been addressed by the scientific community. The most similar work to ours is probably the one published by Guan et al. where the authors propose the automatic hot spot detection and segmentation in PET images. However, not all tumor lesions appear as a hot spot in PET scans, and not all hot spots are related to tumor lesions (e.g., inflammation lesions, thyroids uptake, etc.). Furthermore, its proposed method based on unsupervised learning is highly restricted since it assumes homogeneity of tumoral lesions. Finally, the authors do not provide an exhaustive validation of the proposed system neither at the technological nor at the clinical level in a large set of real PET/CT data.

Several related scenarios have been addressed in this research field, mainly concerted on expert-guided semi-automatic PET tumor segmentation and the application of supervised learning strategies in PET/CT imaging.

Regarding expert-guided semi-automatic PET tumor volume segmentation, a number of works have been done aiming to optimize tumor segmentation results starting from the expert defined initial boundary and predefined algorithm parameters. Alternatives include GraphCuts,[1] spectral clustering,[28] adaptive thresholding,[20] gradient-based methods,[27] fuzzy clustering based methods[3] and iterative thresholding methods.[25]

Machine learning methods, although not aimed at this particular application, have been also applied in PET tumor segmentation. In Ref. [11] the authors apply a learning methodology framework using Support Vector Machines to assist in the threshold-based segmentation of non-small-cell lung cancer tumors in thorax PET/CT imaging for use in radiotherapy planning.

A key point to mention in this supervised learning scenario is the utmost importance of voxel contextual information. That is to say, the tumoral condition of a given voxel is strongly dependent on the characteristics of the surrounding voxels. Several strategies have been introduced in this field, including graphical models,[13, 14, 19, 24] super-pixel methods[4, 7, 9] or contextual priming.[23]

However, most of these previous approaches requires a prestablished prior distribution of spatial properties or the need to consider homogeneous features of neighboring voxels. In our case, tumor lesions cannot be defined to appear spatially distributed with a consistent distribution neither to maintain similar homogeneous properties. Given this fact, we extract a set of heterogeneous multi-modal features and define a contextual learning approach based on discriminative classifiers, which will be able to learn those relevant features and spatial relations present in the training set. For this task, we focus on the Stacked Learning framework[5] and in particular, in the Multi-scale stacked sequential learning (MSSL) alternative presented in Ref. [6]. In that work, the authors extend the Stacked Sequential Learning to include in the feature vector of 2D samples the label predictions in the neighborhood of image pixels at different scales, outperforming state-of-the-art approaches for 2D image segmentation. With the same aim, in this work we present an adapted definition of MSSL in the 3-Dimesional space.

3. MATERIALS

A total of 200 whole-body FDG-PET/CT studies (corresponding to different patients) were obtained from the Philips PET/CT Gemini TF machine located at the nuclear medicine department in the Hospital de Sant Pau (Barcelona, Spain). Half of them (100) correspond to perfectly healthy (control) patients, and the other half (100) correspond to breast cancer patients in some cancer stage, ranging from low to very severe condition (following an approximate uniform distribution, Fig. 2).

Each study contains two co-registered volumes (PET and CT) in DICOM format. From the DICOM metadata, SUV values for PET voxels and HU values for CT voxels can be computed. A PET voxel corresponds to 64 mm^3 and a CT voxel to 2 mm^3. Volume dimensions are $144 \times 144 \times Np$ for PET volumes and $512 \times 512 \times Nc$ for CT volumes (Np and Nc being the number of slices for each volume). Np and Nc are generally related with a ratio of $Nc/Np = 2.66$. However, the actual number of slices is dependent on the volume of interest selected by the acquisition technician, varying with the patient's height and the anatomical limits of interest (typically either from neck to middle-thigh or from the top of the skull to the feet). Patient's position during acquisition is also variable, mainly related to arm positions (Fig. 2).

Ground Truth in PET tumoral volume segmentation for each breast cancer patient was carried out by three independent nuclear medicine experts (*E1, E2, E3*). Mean segmentation overlap between the three datasets, computed using the Jaccard Index,[22] was 0.76 ± 0.07, 0.84 ± 0.04, 0.78 ± 0.05 (*E1 vs. E2, E1 vs. E3, E2 vs. E3*) indicating a high inter-rater reliability. The final Ground Truth dataset was obtained using a majority vote for each voxel from the three raters. Figure 2 shows some ground truth sample segmentations.

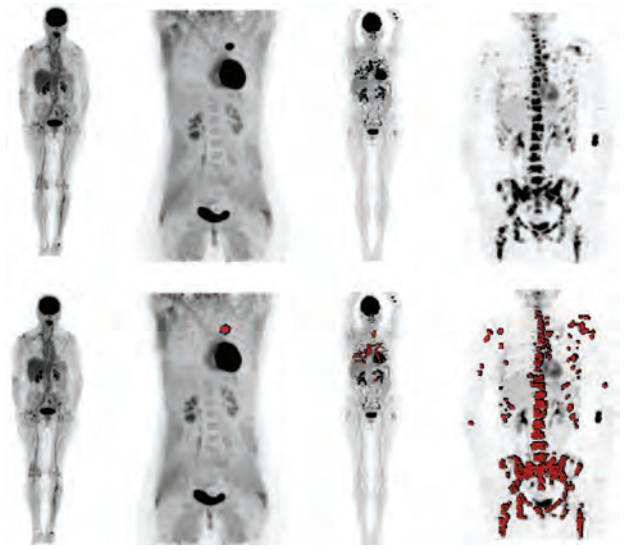


Fig. 2. Sample ground truth PET volumes. Maximum intensity projection is shown for four different patients (top row). Note the high variability in patient's position, number of slices, and pathological condition (healthy, low, middle and high). Ground truth tumoral segmentation for each patient is shown in the bottom row.

4. METHODS

Our supervised learning framework design strategy for multimodal PET 3D tumor segmentation is described as follows. First, a novel voxel feature set is proposed to model the tumoral and physiological conditions within a PET volume. Then, a contextual framework is proposed so that the learning process can take into consideration the 3D context information of a voxel at different spatial scales. Finally, the learning algorithm choice and its implementation are described.

4.1. Voxel Feature Set Proposal

For a PET voxel to be considered tumoral, a set of medical conditions must hold. Thus, clinical knowledge on tumoral volume detection was provided by nuclear medicine experts in order to aid in the feature set design:

(1) Generally, it can be assumed that for a voxel to be tumor-related it must have at least 1.2 in SUV.

(2) Within a whole body PET volume, a big number of voxels with high SUV values corresponding to physiological (i.e., normal) metabolic activity exists. Examples include the heart, the brain, the bladder, the kidneys and the lower gastrointestinal tract. The liver and the thyroid gland may show physiologically moderate uptake but can be pathological if very high SUV focal activity is present. Radiotracer injection point, generally located at the forearm, usually show intense physiological activity.

(3) Less common physiological conditions include muscular and brown fat uptake (showing higher than normal SUV). A hint on detecting these phenomena is the presence of a diffuse and symmetric pattern with respect to the middle-vertical axes (from the neck to the bladder). Hormonal cycles may show variable uptake in breasts and gonads.

(4) The HU value of a voxel given by the CT data is related to the tissue type present at that point. Note that it is not related to the anatomical location or organ where that voxel belongs. For example, a voxel in the biceps, the heart or the calf may have the same HU value (corresponding to the muscle tissue). If the CT scan was obtained using any contrast material,[12] HU values in some anatomical structures may differ from a non-contrast scan.

In order to transform this medical knowledge into quantitative information, a set F of 30 features ($F1, \ldots, F30$) for each PET voxel is described as follows (note that all voxels with less than 1.2SUV value can be discarded for any processing and considered non-tumoral straight away).

Voxel SUV value and mean HU are recorded ($F1$ and $F2$). Mean SUV and HU values on the surrounding of the voxel are

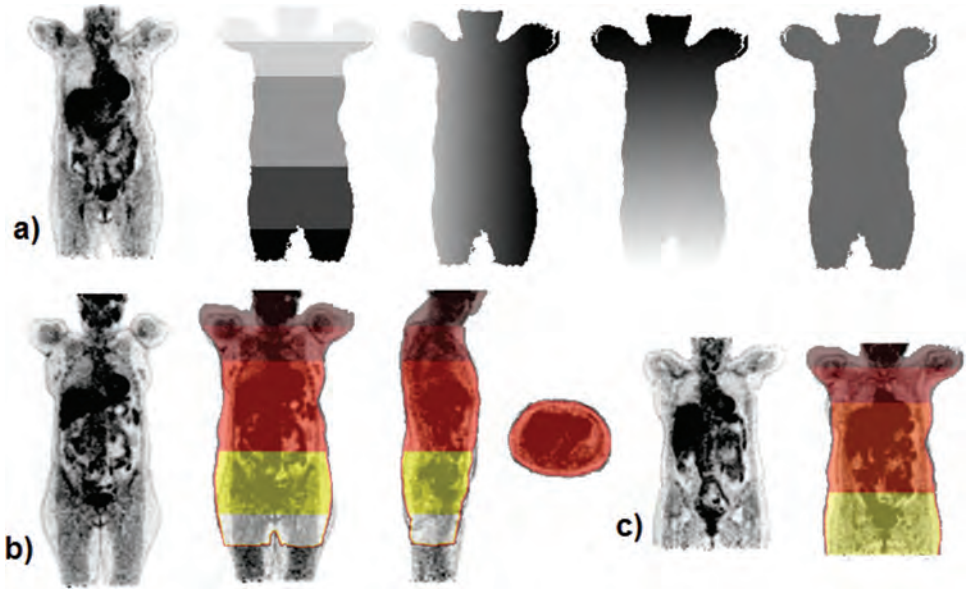


Fig. 3. (a) Sample sagittal slice of the patient selected to build the atlas, anatomical level division and normalized x, y, and z coordinates of the same atlas slice. (b) Sample sagittal slice of a test patient with the fitted atlas overlaid on it (sagittal, coronal and axial views) and (c) a smaller patient atlas fit (note the reslicing in the coregistration process).

also an interesting value for metabolic and anatomical information, so they are recorded for cubical contexts centered at the voxel of sides 3, 5, 9 and 15 (which correspond to cubic regions of about 0.1, 1, 2 and 3 cm^3 side within the patient's body); obtaining a total of 8 new features ($F3, \ldots F10$).

Anatomical location of the voxel within the patient's body is also an important variable. Voxel coordinates $c = (x, y, z)$ of the volume model are not a good choice since PET volumes have variable number of slices, orientation, and include a different portion of the patient's body. Thus, a set of normalized coordinates are needed. To deal with this problem, an "anatomical origin of coordinates" is required. Every voxel of a patient's PET volume should be indexed to a common anatomical reference point. A stable choice for this "anatomical origin of coordinates" is the lower point of the patient's bladder. Bladder is a medium sized organ that shows a very high SUV value in PET scans. Brain and heart could be alternatives, but the brain is not always present the PET scan and the heart does not always show high uptake (see Fig. 2).

An ad-hoc algorithm for bladder location was developed. It is based on finding the most down-centered medium-sized connected component with very high SUV value (75% of maximum SUV in the volume). That connected component has a very high chance of being the patient's bladder. The lower (minimum z coordinate) point of this connected component is defined as the patient's anatomical origin of coordinates $A_o = (a_1, a_2, a_3)$. This method has achieved over 99% accuracy, failing in only 3 cases out of 450 PET test volumes, which were later discarded from further processing.

Although bladder anatomy and physiology show differences among patients, the authors consider this origin of coordinates a robust approximation to a common anatomical location of any PET volume (Fig. 4). Therefore, any voxel coordinates should be centered with respect to A_o, giving the new coordinates $c' = (a_1 - x, a_2 - y, a_3 - z)$. Also, in order to normalize distance between points in very different anatomies, the coordinates should be normalized by the patient's body surface area *bsa* (Mosteller). Thus, the final normalized coordinates of the voxel are given by $C_N = c'/bsa$, which contribute to 3 new features in the voxel's feature set ($F11, F12, F13$).

Once a set of normalized coordinates within the patient's anatomy have been computed, a set of global coordinates are proposed. Its main goal is to give some insight into questions such as "is the voxel near the head? Is it near the thorax? Is it near the lower gastrointestinal tract? Is it in any patient's leg? etc."

A common approach to obtain this kind of information is to fit an atlas to the volume under study. Therefore, an ad-hoc whole body PET atlas was build based on an average sized patient in the most common acquisition position. Since whole body anatomy is extremely variable, the only labels that were included in the atlas were: head and neck, from neck to heart, from the heart to the kidneys level, from the kidneys level to the bladder and from the bladder to the knees. Continuous atlas information was also recorded for each atlas voxel in the form normalized 3D coordinates (within the $[0, 1]$ range) with respect to the upper left vertex of the atlas minimum bounding box. This proposed atlas is then registered to the given PET volume (Fig. 3) using normalized mutual information,[18] obtaining four new features for each PET voxel ($F14, F15, F16, F17$).

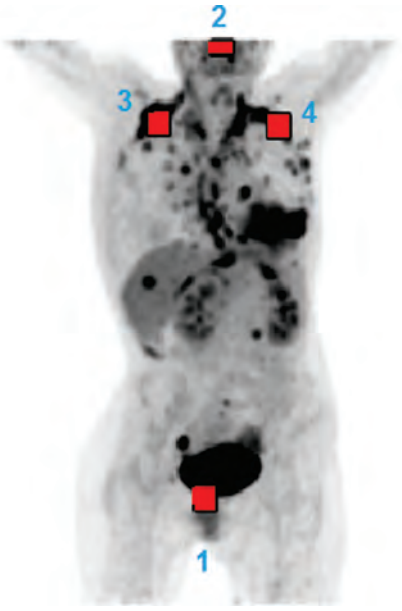


Fig. 4. Sample PET scans of a patient with brown fat uptake. Anatomical origin or coordinates (A_0) (1), middle-head point (2), sample voxel (3) and its symmetric counterpart (4) with respect to the sagittal plane given by (1) and (2). Note that the high symmetry measure between (3) and (4) is able to model the underlying brown fat (i.e., non-tumoral) uptake ocurring those regions.

Another common characteristic of some type of tumor lesions in PET scans is its focal activity, related with a metabolic "hot spot"[2] appearance within the volume. In order to try to include this phenomena for each voxel as a feature, the size of the connected component where a voxel belongs to within its thresholded PET volume at 1.2SUV was recorded (*F18*).

In order to define a symmetric metric, the symmetry plane must be defined and computed. Within this medical context, symmetric muscular or brown fat uptake is defined with respect to the central sagittal plane of a patient's body. Given that the plane orientation is known (sagittal), two central points are needed to completely define it. One of them is the already computed A_o, approximately located at the low-center of a patient's body. The other could be the head mid-point, easily computed taking the average of the middle points of the central connected component of the first slices of any whole-body CT volume (Fig. 4). Now, the symmetry plane equation can be computed and, for each voxel, its symmetric voxel with respect to that plane can be located. Finally, a quantitative measure on the symmetric counterpart of any voxel activity is obtained computing subtracting the mean SUV value on a $3 \times 3 \times 3$ cube centered at that voxel and the corresponding value on the same cube type centered at the voxel's symmetric counterpart (*F19*).

Recording local heterogeneity information is also useful in a segmentation learning framework. Thus, the 3D gradient vector and its magnitude are computed for any PET and CT voxel ($F20, \ldots, F27$).

Binary information regarding patient's sex and the use of contrast material in the CT scan is also included, which is found in the DICOM metadata ($F28, F29$).

Finally, the patient's arms position at the time of the scan (upwards or downwards) is derived using a simple ad-hoc algorithm based on counting the number of connected components in a thresholded, smoothed and dilated version of the top scan

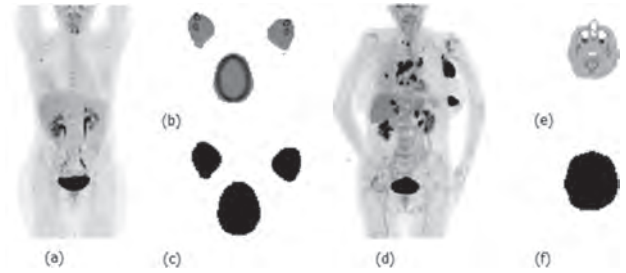


Fig. 5. Arm position detector. MIP projections (a), (d), one of the patient's firsts resized axial CT slice thresholded (b), (e) and a smoothed, dilated and thresholded version of it (c), (f). Counting the number of connected components in these last images indicates whether the patient's has got its arms upwards or downwards.

slices: if there is only one component (the head), the patient has its arms down and it there are three components (arms and head) she has its arms up (Fig. 5). This method has achieved over 99% accuracy, failing in only 6 cases out of 450 PET test volumes. This last binary feature (*F30*) is relevant since the radiotracer injection point, located at a forearm, is an important potential false-positive source of error, and the learning system should know whether that source is located at the head level or at the hip level.

The summarized list of features computed for each voxel from the different image modalities are summarized in Table I.

4.2. A 3D Contextual Learning Framework

Once all voxel features are computed for each PET scan in the data set, the training matrix X (of size 15719349×30 in our case) can be obtained and a base classifier model C can be trained with any supervised learning algorithm h using the voxel's Ground Truth labels (g):

$$C = h(X, g)$$

Then, this classifier can be tested on the dataset, obtaining for each voxel a predicted binary classification label y on its tumoral condition. As it has been mentioned before, if a voxel has been classified as tumoral, its neighborhood voxels are more likely to be classified as such than the others. The goal of the proposed 3D Multiscale Stacked Sequential Learning (MSSL) is to augment the feature set of each voxel with contextual information from the base classifier predictions at neighboring voxels in order to train a new classifier that would be able to learn the application's context rules and outperform its global segmentation results.

In particular, the original feature set will be extended with a set of contextual parameters (Z) using the base classifier predictions (y) in several 3D context scales (S). Then, a new classifier C' will be trained using the extended training matrix X' obtaining the final contextual classifier. This approach is summarized in Figure 6.

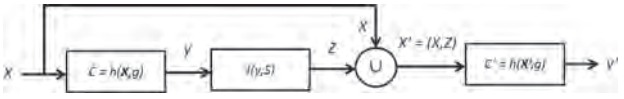


Fig. 6. 3DMSSL approach in building a contextual classifier (C') extending the original feature set X with a set of contextual parameters (Z) computed using a context function (J) that uses the prediction labels (y) of a base classifier (C) in different context scales (S) for each voxel.

In order to deal with this task, the first step is to define appropriate neighborhood dimensions and structure (S) for each voxel. Cubic neighborhoods of size 3, 5, 9 and 15 voxels have empirically shown good performance results. It could be stated that Gaussian neighborhoods should be more suitable for this application, since human anatomy does not contain straight borders. However, using cubic boundaries would facilitate the incorporation of parallel computation techniques in this computing-intensive scenario.

Once the neighborhood scales are defined, the contextual information to be recorded for each neighborhood scale should be defined (J). The following set of contextual parameters (Z) for each scale, based on the information given by the predicted labels y of the base classifier (C), is proposed:

$$Z1_S = \frac{\#\text{voxels in scale S classified as tumoral by C}}{\#\text{total voxels in S}}$$

$$Z2_S = \frac{\#\text{voxels in scale S classified as nontumoral by C}}{\#\text{total voxels in S}}$$

$$Z3_S = \frac{\#\text{voxels in scale S not included in the training process of C}}{\#\text{total voxels in S}}$$

Note that $Z3$ is included since all voxels with less than 1.2 SUV are discarded for the learning process. Then, the original feature set for each voxel is extended with these contextual parameters:

$$F' = FU\{Z1_s, Z2_s, Z3_s\}\,\forall S$$

which in practice corresponds to appending 12 features (4 scales, 3 parameters) to each original voxel's feature set ($F31, \ldots, F42$) obtaining an extended training matrix (X'). Finally, the contextual classifier C' is trained on the new training data X':

$$C' = h(X', g)$$

which will be used as the final classifier system for our automatic whole body PET tumor segmentation system proposal.

4.3. Learning Algorithm Choice and Implementation

Finally, the supervised learning algorithm (h) choice and implementation will be discussed.

Table I. Voxel feature set design summary.

Feature groups	Description
$F1 \ldots F10$	SUV and HU values at the voxel and its neighborhood lattices.
$F11 \ldots F13$	Voxel's local anatomically normalized coordinates.
$F14 \ldots F17$	Voxel's global anatomical location approximation.
$F18 \ldots F19$	Voxel's hot spot belonging and symmetry information.
$F20 \ldots F27$	Voxel's PET and CT volume gradient information.
$F28 \ldots F30$	Binary information regarding the patient's sex, CT type and arms position.

Given the large amount of data to be processed within the learning framework (nearly half a million voxels per patient to be processed), a cascading strategy is required. Since the training set is large, there is a big number of training patterns and we are faced with a two-class problem (a voxel is either tumoral or not), the authors have chosen a Cascade of classifiers as the learning framework.[30] The learning algorithm chosen as building block in the cascade of classifiers is the Discrete Adaboost classifier with decision stumps. The selection of this alternative classifier is two-fold. First, it performs feature selection within the learning process, which allows studying the discriminative power of each feature in the classification rule, which is a desirable behavior given the clinical context. And second, it does not require any hyper parameter tuning a part from the number of iterations, which can be set sufficiently high in the first place, resulting in a faster training stage in comparison with other state of the art approaches.

The final implementation scheme is shown in Figure 7. Given the large number of non-tumoral training entries, a single AdaBoost classifier cannot be trained using the whole dataset due to computational resource limitations. Instead, the training set is initially divided in the set of tumoral entries T (with a total of 190537 entries) and a subset of the non-tumoral entries with the same size as $T(H_0)$. A first classifier C_1 is trained with this data using an AdaBoost learning block. After testing this classifier with its training entries, true negatives instances are rejected from further processing since they have been correctly classified by this cascade level. False positive, false negative and true positive instances are pipelined to the next block level in conjunction with a new subset of non-tumoral entries that are misclassified by the previous levels (H_{FN_1}) in order to train the second level classifier (C_2). This process is repeated until all non-tumoral entries are considered in any level. To avoid noise overfitting, it is recommended to discard a small fraction of false positive entries in a subset of cascade levels.

Once the cascade of AdaBoost classifiers has been trained, the testing procedure for any given voxel entry is accomplished in the following manner: if classifier C_1 predicts it as a non-tumoral voxel, it is finally classified as such without further processing. Otherwise, the entry is tested on the next level classifier (C_2). This procedure is repeated until the entry is either classified as non-tumoral by any cascade level or the end of the cascade is reached, in which case the entry is classified as tumoral.

For each cascade level C_i, the AdaBoost classification parameters are computed, which include a set of feature weights that are used to compute the voxel's classification value, and a threshold

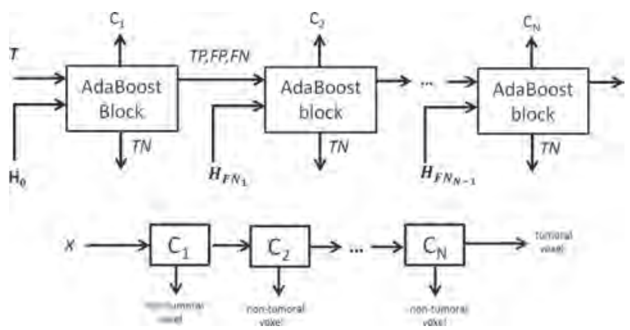


Fig. 7. Supervised learning framework a cascade of adaboost classifiers. Training (top) and testing (bottom) phases.

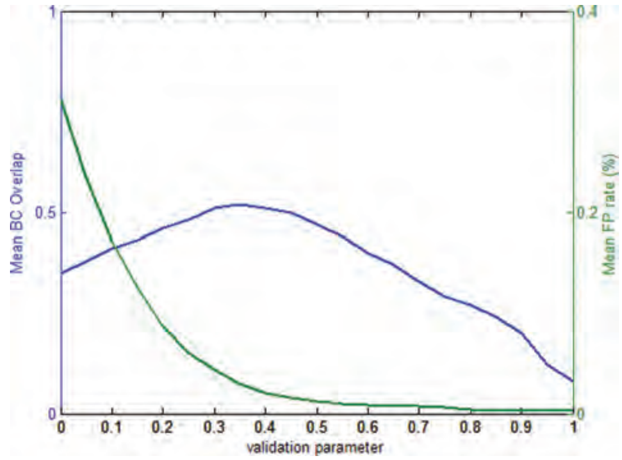


Fig. 8. Sample validation process. Without loss of generality, when the validation parameter (p) is varied from 0 to 1, a decreasing in mean FP rate and an increasing in mean BC overlap of the validation set (up to a local maximum) is observed.

that will be applied to it to obtain the final voxel class (either tumoral or non-tumoral). Since on average there is only one tumoral voxel entry for each 120 non-tumoral voxel entries, the training procedure is biased to produce a high false positive rate.

To overcome this limitation, a validation stage is designed as follows. For each classifier, the difference between the mean tumoral and non-tumoral voxel classification values is computed (D_L). Then, a fraction p of D_L is added to the classifier's threshold in order to increase its strictness and decrease its false positive rate. When performing a p sweep in the range 0 to 1 and checking the classifier's performance in a separate validation set, a local maximum in mean Jaccard overlap Index in the set of breast cancer PET volumes is always observed (Fig. 8). Therefore, the optimal validation parameter $p_{\max}$ is recorded and all cascade classifier thresholds are modified with an additive term $p_{\max} \cdot D_L$.

5. RESULTS AND DISCUSSION

In this section, the proposed system results are described both at the technological and clinical level. First, performance results of the learning framework are presented. Then, a comparison with a justified alternative to the proposed system is shown and the clinical relevance of the system is illustrated.

Performance of the proposed learning scheme is addressed using a 10-fold cross-validation approach in the test phase. From the 200 patient's PET/CT studies original dataset, for each learning round, 10% (20 patients, 10 breast cancer patients and 10 control patients) are used as a test set and 90% (180 patients) as training and validation set, of which 5% (9 patients) are used as validation set and the remaining 171 patients as training set.

Performance results in the test set are computed using the mean overlap (mOV) metric for segmentation accuracy and the mean sensitivity and specificity for generic classifier accuracy. The overlap metric (based on the Jaccard Index) is particularly suitable for this application since it is and standard and rigorous evaluation of volume segmentation that penalizes both under and over segmentation estimations.

Note that for any GT control patient, its overlap parameter with the classifier output is either 0 or 1, being 1 only when the

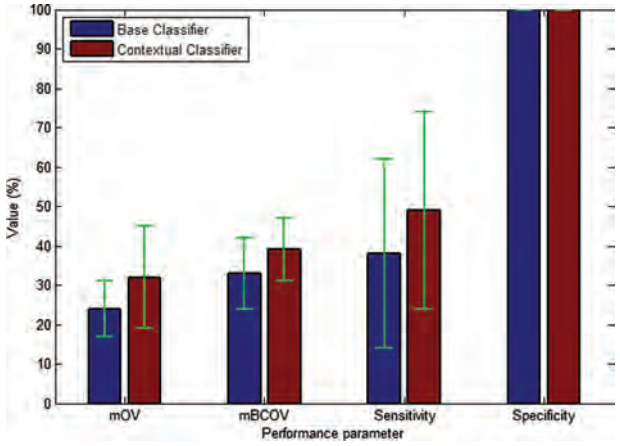


Fig. 9. Global performance results of the base and contextual classifiers.

outputs classifier has no false positive voxel; a single classifier's false positive voxel on a control patient would give 0 overlap. Since half of the test patients are control patients, this phenomena will alter the mean overlap metric in an unstable manner. A more convenient mean overlap performance is obtained if overlap is only computed on breast cancer patients (mBCOV). Then, the performance on the control patients can be analyzed using the sensitivity and specificity parameters.

Performance results at the voxel level obtained by the base (C) and contextual (C') classifiers are shown in Tables II and III. Figure 9 summarizes the global classifiers' test performance.

Note the significant improvement of the contextual classifier with respect to the base classifier (nearly 20% in mean overlap, paired t-test $p < 0.03$). The higher specificity than sensitivity result is coherent with the average presence of one voxel tumoral entry for each 120 non-tumoral entries during the learning process.

The combination of a high mean overlap (39%) achievement for this scenario and nearly mean 50% sensitivity points out that the learning system is actually detecting most of tumor lesions but doing a very strict segmentation of them, classifying a high percentage of its boundary voxels as false negative. However, the authors consider that this problem could be addressed with appropriate post processing techniques (such as region-growing or 3D morphology) aiming to obtain much better overall results. It could be stated that the proposed learning framework principal's target is to detect any tumoral region within a whole-body PET volume, and successful results in this respect have been shown.

Table II. 10-fold cross-validation base classifier (C) results.

Round	mOV (%)	mBCOV (%)	Sensitivity (%)	Specificity (%)
1	23.00	46.00	80.84	99.99269
2	25.15	40.29	60.12	99.99781
3	14.01	28.02	30.47	99.98914
4	37.98	15.97	8.72	99.99896
5	20.02	30.05	16.67	99.99608
6	24.51	39.02	46.99	99.99709
7	32.84	25.69	6.79	99.99787
8	13.28	26.57	33.19	99.99453
9	28.88	37.76	56.45	99.99505
10	21.43	42.85	47.80	99.99630
Mean	**24 ± 7**	**33 ± 9**	**38 ± 24**	**99.995 ± 0.003**

Table III. 10-fold cross-validation contextual classifier (C') results.

Round	mOV (%)	mBCOV (%)	Sensitivity (%)	Specificity (%)
1	35.80	51.60	94.13	99.98736
2	21.95	43.91	81.26	99.99270
3	13.44	26.88	41.23	99.98801
4	47.53	45.05	42.99	99.98946
5	30.29	40.58	20.02	99.99728
6	50.19	30.38	40.90	99.99834
7	47.75	35.49	27.41	99.99611
8	13.94	27.87	26.65	99.99669
9	24.06	48.12	75.90	99.99033
10	40.99	41.99	38.92	99.99844
Mean	**32 ± 13**	**39 ± 8**	**49 ± 25**	**99.993 ± 0.004**

Several learning process performance results are worth mentioning. Figure 10 shows the classifier's performance evolution at each level of the cascade process for the base and contextual classifiers considering the mean of training, validation, and test data set performances. Note that the contextual classifier needed much less number of cascade levels (10) with respect to the base classifier (21) for achieving better segmentation results, which indicates that the contextual information provided by the base classifier contributed significantly in the increase of overall learning framework performance.

In order to analyze the relevance of each selected feature during the learning process, Figure 11 shows the mean relative feature weight values estimated using the set of AdaBoost confidence parameters computed at each cascade level in the learning process.

The most relevant features from the base feature set are the mean SUV and HU values in the lowest scale, fitted atlas coordinates and PET and HU gradient magnitudes. These results are supported at clinical level since these features are the most considered by nuclear medicine physicians in clinical PET tumoral volume detection. Not surprisingly, less relevant features are related to global binary information such as the patient's gender, CT type and arms position. Contextual features relevance during the contextual learning process is high and notably variable across scales (S1, S2, S3, S4) and parameters, which is coherent with the classifier's effort to learn the complex whole body human anatomy and physiology in different scales.

Visual classifiers' results to be compared with Figures 1–2 ground truth segmentation are shown in Figure 12, Where one can see how the proposed systems obtains a good approximation of the real tumoral volume segmentation. Also note the contextual classifier outperformance with respect to the base classifier in most tumor lesion segmentations.

Using a fully-MALTAB implementation of the proposed learning framework, computation time for the training phase was about 28 hours. Given a test whole body PET volume, the total computation time including voxel feature computation and its posterior classification is about 7 minutes in a 64 bit Intel(R) Core™ i5 CPU (750@2.67 GHz 2.67 GHz) with 8 GB RAM and Windows 7 OS.

Regarding the clinical impact of the proposed tool, note that the clinical use of the obtained whole body metabolic tumor volume segmentation relies on the a posteriori computation of a set of prognostic parameters from it.[21,29] A common prognostic parameter that have proven useful in cancer management is precisely the total whole body metabolic tumor volume (WBMTV),

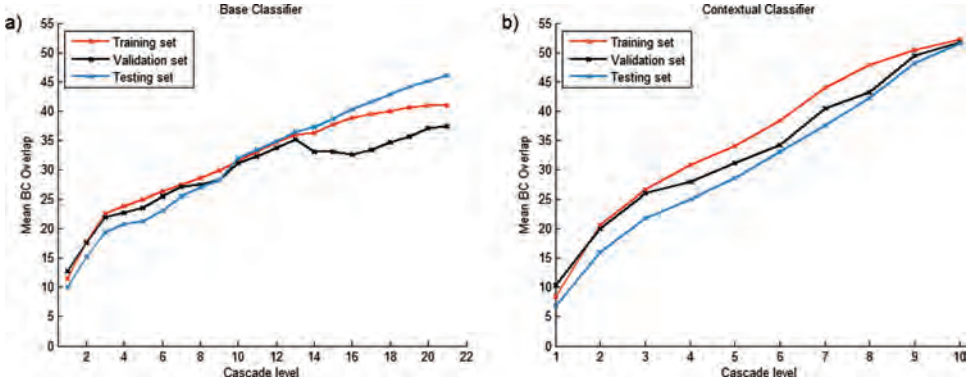


Fig. 10. Base (a) and contextual (b) classifiers' performance evolution at each cascade level of the learning process.

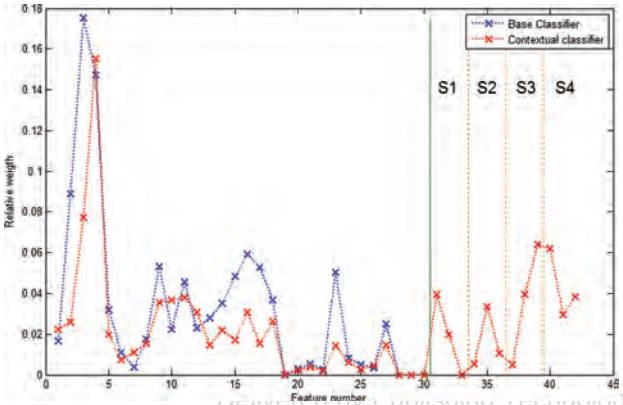


Fig. 11. Relative feature weights. Base classifier (blue, first level of the learning process), contextual classifier (red, second level of the learning process with the extended contextual feature set).

as shown in Ref. [17], trivially computed counting the total number of tumoral voxels of the PET volume and converting to volume units such as cm^3.

Therefore, a relevant clinical performance parameter is the correlation obtained at comparing the WBMTV values obtained from the expert-guided tumor segmentation masks and the proposed automatic segmentation framework. A Pearson correlation coefficient of 73% is obtained.

Finally, a comparative analysis of our proposed system with another completely automatic approach to WBMTV computation is carried out. A rather naïve alternative is based on a direct thresholding masking of the PET volume. Clinically, it has been stated that although it is variable according to tumors, a 3.0 threshold on SUV value is a general cut-off set for differentiating between malignant and benign lesions.[17] Thus, direct thresholding on 3.0 SUV value can be used as a naïf automatic whole body metabolic tumor volume segmentation system. Note however the

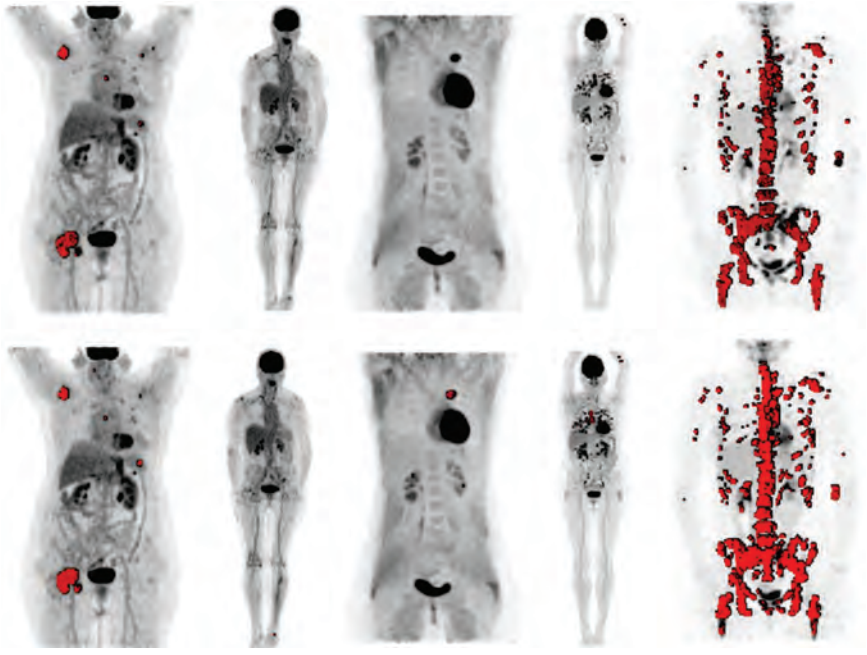


Fig. 12. Base classifier results on patients of Figures 1–2 (first row), contextual classifier results (second row). Ground truth data for these samples is shown in Figures 1 and 2, respectively.

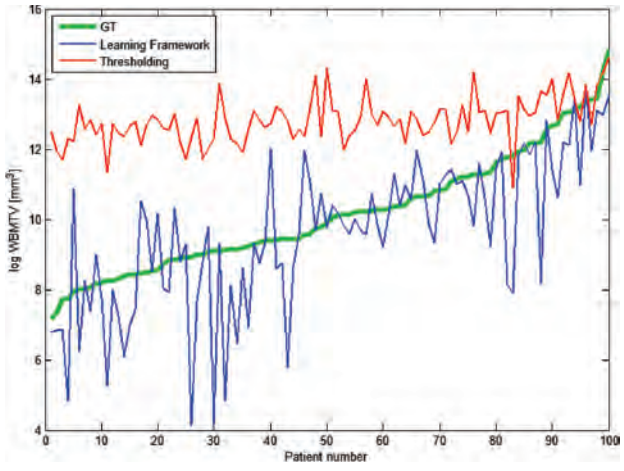


Fig. 13. Correlation between GT WBMTV and its automatic approximations using the proposed supervised learning framework (73%) or direct thresholding in 3.0 SUV value (64%). Note that only breast cancer patients are included and patient numbers are sorted by GT WBMTV in ascending order.

big limitation of this methodology, which will considerer any physiological uptake higher than 3.0 (brain, heart, bladder, kidneys, etc.) as tumoral volume. Performance results of this alternative show notably poorer results ($5 \pm 2\%$ mean overlap, $10 \pm 4\%$ mean BC overlap, $72 \pm 18\%$ sensitivity and $99.86 \pm 0.03\%$ specificity). Correlation with the WBMTV parameter calculation also shows worst results (64%), as shown in Figure 13.

6. CONCLUSIONS AND FUTURE WORK

In this paper, a supervised learning framework has been proposed for solving the whole-body breast cancer PET/CT metabolic tumor volume segmentation problem. Our approach is based on training a cascade of AdaBoost classifiers and a 3D contextual learning framework from a set of automatically computed multi-modal PET/CT features. Given the complexity of the addressed problem, system's performance has shown good results at the technological level (49% sensitivity, over 99.99% specificity and 39% overlap) and at the clinical level (73% correlation in metabolic tumor volume calculation with respect to medical experts).

Future work for this project include extending the training dataset to increase performance results, testing the framework on different cancer types in order to try to obtain a general PET/CT metabolic tumor volume segmentation tool (which would be valuable in nuclear medicine departments), the proposal of post processing techniques to improve the overall segmentation accuracy, improving computation time using parallel and GPU techniques in a C++ environment and checking for online learning alternatives. An incorporation protocol of the proposed system within the clinical scenario is also required, which will require its integration with the hospital's picture archiving and communication system (PACS), the implementation of a user-friendly graphical user interface and the incorporation of a semi-automatic segmentation module to allow the medical expert to correct the system's automatic segmentation proposals.

Acknowledgments: The work of Frederic Sampedro is supported by the Spanish government FPU (Formación del Profesorado Universitario) doctoral grant.

References and Notes

1. C. Ballangan, X. Wang, M. Fulham, S. Eberl, and D. Dagan, Lung tumor segmentation in PET images using graph cuts. *Computer Methods and Programs in Biomedicine* 109, 260 **(2013)**.
2. O. Bauwens, M. Dusart, P. Pierard, J. Faber, T. Prigogine, B. Duysinx, B. Nguyen, M. Paesmans, J. Sculier, and V. Ninane, Endobronchial ultrasound and value of PET for prediction of pathological results of mediastinal hot spots in lung cancer patients. *Lung Cancer* 61, 356 **(2008)**.
3. S. Belhassen, C. Llina, and H. Zaidi, A new fuzzy clustering-based segmentation of heterogeneous 18F-FET PET tumors for definition of gross target volume in high-grade glioma. *NeuroImage* 47, 39 **(2009)**.
4. T. Cour and J. Shi, Recognizing objects by piecing together the segmentation puzzle, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition IEEE Computer Society*, Minneapolis, Minnesota, USA **(2007)**, pp. 1–8.
5. T. Dietterich, Machine learning for sequential data: A review, *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, Springer-Verlag, London, UK **(2002)**, pp. 15–30.
6. C. Gatta, E. Puertas, and O. Pujol, Multi-scale stacked sequential learning. *Pattern Recognition* 44, 2414 **(2011)**.
7. L. Gorelick and R. Basri, Shape based detection and top-down delineation using image segments, *Int. J. Comput. Vision* 83, 211 **(2009)**.
8. H. Guan, T. Kubota, Huang, X. Sean, X. S., and M. Turk, Automatic hot spot detection and segmentation in whole body FDG-PET images. *Image Processing, IEEE International Conference*, October **(2006)**, p. 85.
9. G. Heitz and D. Koller, Learning spatial context: Using stuff to find things, *Proceedings of the 10th European Conference on Computer Vision*, Springer-Verlag, Marseille, France **(2008)**, pp. 30–43.
10. M. Hoffer, CT Teaching Manual, Georg ThiemeVerlag **(2000)**, p. 12.
11. A. Kerhet, C. Small, H. Quon, T. Riauka, L. Schrader, R. Greiner, D. Yee, A. McEwan, and W. Roa, Application of machine learning methodology for PET-based definition of lung cancer. *Curr. On. Col.* 17, 41 **(2010)**.
12. K. Kitajima, Y. Ueno, K. Suzuki, M. Kita, Y. Ebina, H. Yamada, M. Senda, T. Maeda, and K. Sugimura, Low-dose non-enhanced CT versus full-dose contrast-enhanced CT in integrated PET/CT scans for diagnosing ovarian cancer recurrence, *European Journal of Radiology* 81, 3557 **(2012)**.
13. S. Kumar and M. Hebert, Discriminative random fields: A discriminative framework for contextual interaction in classification, *Proceedings of the 2003 IEEE International Conference on Computer Vision (ICCV '03)*, Nice, France **(2003)**, Vol. 2, pp. 1150–1157.
14. J. D. Lafferty, A. McCallum, and F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proceedings of the International Conference on Machine Learning*, Williamstown, MA, USA **(2001)**, pp. 282–289.
15. G. Lucignani, G. Paganelli, and E. Bombardieri, The use of standardized uptake values for assessing FDG uptake with PET in oncology: A clinical perspective. *Nuclear Medicine Communications* 25, 651 **(2004)**.
16. R. D. Mosteller, Simplified calculation of body-surface area. *N. Engl. J. Med.* 317, 1098 **(1987)**.
17. J. Oh, J. Seo, A. Chong, J. Min, H. Song, Y. Kim, and H. Bom, Whole-body metabolic tumour volume of ^{18}F-FDG PET/CT improves the prediction of prognosis in small cell lung cancer. *European Journal of Nuclear Medicine and Molecular Imaging* 39, 925 **(2012)**.
18. J. Pluim, A. Maintz, and M. Viergever, Mutual information based registration of medicalimages: A survey. *Transactions on Medical Imaging* 22, 986 **(2003)**.
19. X. Richard, R. Zemel, and M. Carreira, Multiscale conditional random fields for image labeling, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washing-ton, DC, USA **(2004)**, pp. 695–702.
20. D. Schinagl, W. Vogel, A. Hoffmann, J. V. Dalen, W. Oyen, and J. Kaanders, Comparison of five segmentation tools for 18F-fluoro-deoxy-glucose-positron emission tomography-based target volume definition in head and neck cancer. *International Journal of Radiation Oncology*Biology*Physics* 69, 1282 **(2007)**.
21. A. Takeda, N. Yokosuka, T. Ohashi, E. Kunieda, H. Fujii, Y. Aoki, N. Sanuki, N. Koike, and Y. Ozawa, The maximum standardized uptake value (SUVmax) on FDG-PET is a strong predictor of local recurrence for localized non-small-cell lung cancer after stereotactic body radiotherapy (SBRT). *Radiotheraphy and Oncology* 101, 291 **(2011)**.
22. P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, ISBN 0-321-32136-7 **(2005)**.
23. A. Torralba, Contextual priming for object detection, *Int. J. Comput. Vision* 53, 169 **(2003)**.
24. S. Vishwanathan, N. Schraudolph, M. Schmidt, and K. Murphy, Accelerated training of conditional random fields with stochastic gradient methods. *Proceedings of the 23rd International Conference on Machine Learning, ACM*, New York, NY, USA **(2006)** pp. 969–976.
25. W. Jentzen, L. Freudenberg, E. Eising, M. Heinze, W. Brandau, and A. Bockisch, Segmentation of PET Volumes by Iterative Image Thresholding. *J. Nucl. Med.* 48, 108 **(2007)**.

26. W. Wang, H. Jun, D. Wang, and Y. Yin, A comparison of FDG PET-CT tumor segmentation for clinical application. *Applied Mechanics and Materials* 195, 572 **(2012)**.
27. M. Werner, A. Nelson, W. Choi, Y. Arai, P. Faulhaber, P. Kang, F. Almeida, Y. Xiao, N. Ohri, K. Brockway, J. Piper, and A. Nelson, What is the best way to contour lung tumors on pet scans? multiobserver validation of a gradient-based method using a NSCLC digital PET phantom. *International Journal of Radiation Oncology*Biology*Physics* 82, 1164 **(2012)**.
28. F. Yang and P. Grigsby, Delineation of FDG-PET tumors from heterogeneous background using spectral clustering. *European Journal of Radiology* 81, 3535 **(2012)**.
29. H. Zhang, K. Wroblewski, S. Liao, R. Kampalath, B. Penney, Y. Zhang, and Y. Pu, Prognostic value of metabolic tumor burden from ^{18}F-FDG PET in surgical patients with non-small-cell lung cancer. *Academic Radiology* 20, 32 **(2013)**.
30. H. Zhaofeng, T. Tieniu, and S. Zhenan, Topology modeling for adaboost-cascade based object detection. *Pattern Recognition Letters* 31, 912 **(2010)**.

Received: 22 December 2013. Accepted: 25 June 2014.

Original article

Nuclear Medicine Communications

Obtaining quantitative global tumoral state indicators based on whole-body PET/CT scans: a breast cancer case study

Frederic Sampedro[a], Anna Domenech[c] and Sergio Escalera[b,d]

Objectives **In this work we address the need for the computation of quantitative global tumoral state indicators from oncological whole-body PET/computed tomography scans. The combination of such indicators with other oncological information such as tumor markers or biopsy results would prove useful in oncological decision-making scenarios.**

Materials and methods **From an ordering of 100 breast cancer patients on the basis of oncological state through visual analysis by a consensus of nuclear medicine specialists, a set of numerical indicators computed from image analysis of the PET/computed tomography scan is presented, which attempts to summarize a patient's oncological state in a quantitative manner taking into consideration the total tumor volume, aggressiveness, and spread.**

Results **Results obtained by comparative analysis of the proposed indicators with respect to the experts' evaluation show up to 87% Pearson's correlation coefficient when providing expert-guided PET metabolic tumor volume segmentation and 64% correlation when using completely automatic image analysis techniques.**

Conclusion **Global quantitative tumor information obtained by whole-body PET/CT image analysis can prove useful in clinical nuclear medicine settings and oncological decision-making scenarios. The completely automatic computation of such indicators would improve its impact as time efficiency and specialist independence would be achieved.** ***Nucl Med Commun*** **35:362–371 © 2014 Wolters Kluwer Health | Lippincott Williams & Wilkins.**

Nuclear Medicine Communications 2014, **35**:362–371

Keywords: image analysis, oncology, PET, quantification

[a]Faculty of Medicine, [b]Computer Vision Center, Autonomous University of Barcelona, [c]Nuclear Medicine Department, Hospital de Sant Pau and [d]Faculty of Mathematics, University of Barcelona, Barcelona, Spain

Correspondence to Frederic Sampedro, MSc, Faculty of Medicine, Autonomous University of Barcelona, Barcelona 08193, Spain
Tel: +34 699 805 231; fax: +34 934 037 151;
e-mail: fredsampedro@gmail.com

Received 21 October 2013 **Revised** 20 November 2013
Accepted 25 November 2013

Introduction and related work

^{18}F-fluorodeoxyglucose (^{18}F-FDG) PET/computed tomography (PET/CT) has become a standard imaging method for the staging, restaging, and monitoring of treatment response in a variety of tumors. By injecting the ^{18}F-FDG radiopharmaceutical into the patient, a metabolic image of the whole body, measured in standard uptake value (SUV) units, is acquired. This metabolic image is obtained in combination with a coregistered CT scan that provides higher anatomical resolution (in Hounsfield units, HU).

Whole-body (WB) PET/CT scans are a valuable tool for cancer detection and can be used to evaluate the spread of cancer throughout the patient's body [1,2]. The current analysis of WB PET/CT scans is mainly visual; nuclear medicine physicians build a descriptive report about their findings regarding the possible location of cancer and its metastases.

Local quantitative tumor lesion information, such as its mean and maximum uptake value (SUV_{mean}, SUV_{max}) and diameter, is usually included in the report. Global quantitative information, such as the whole-body metabolic tumor volume (WBMTV) and total lesion glycolysis (TLG), is usually not included in the report, although they have been proven to be clinically relevant as independent prognostic markers [3–7]. This may be partly because the measurement of these parameters, which currently requires an expert-guided manual or semiautomatic tumor segmentation from the PET scan, is highly time consuming and therefore not practical in a clinical setting. It may also be because the usefulness of this time-inefficient measurement has not been fully determined [7].

In this work we address the computation of global quantitative indicators from WB PET/CT scans that reflect the patient's oncological state. This type of indicator is referred to as PET Global Oncological State Indicator (PGOSI) hereon. Here, we consider that the oncological state of a patient is deduced in a qualitative manner from the expert-based visual analysis of WB PET images and is related to the quantity of tumor present in the body as well as its aggressiveness and spread. Clinical nuclear medicine experts agree on the need for such a quantitative indicator that, when combined with complementary oncological indicators such as tumor markers

0143-3636 © 2014 Wolters Kluwer Health | Lippincott Williams & Wilkins DOI: 10.1097/MNM.0000000000000067

or biopsy results, would prove a valuable tool for oncological decision making [1,2]. WBMTV and TLG can be considered examples of PGOSI. Some conceptual limitations of these indicators as well as new indicator proposals that try to overcome them are presented in the subsequent sections.

In this scenario, defining a gold standard for assessing the performance of any proposed PGOSI is a complex task. It could be argued that an appropriate choice would be to compare the PGOSI results with other oncological clinical variables such as biopsy results, TNM staging [8], tumor markers, or *N*-year survival rates. However, we consider that none of these variables appropriately model what our PGOSI proposal is intended for: biopsy results are only conclusive about a single anatomical location and do not relate directly to the total tumor quantity and spread within the patient's whole body; TNM staging does give an insight into the tumor quantity and spread but in a categorical manner, and hence it could be argued that two patients may possess slightly different oncological states albeit belonging to the same TNM category; tumor marker results may be independent of PET/CT observations depending on the type of tumor and its stage, and *N*-year survival rates may not be appropriate for comparison with PGOSI results as patients may undergo different treatments and suffer from other nononcological pathologies.

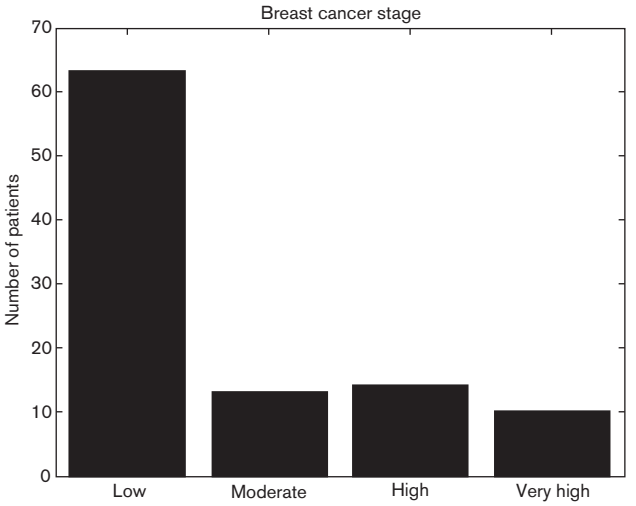
Note that in the related scenario of PET follow-up evaluation, in which two time consecutive PET/CT scans are compared to address therapy response, treatment outcome parameters can be successfully used as the gold standard to address the performance of the proposed quantitative indicators obtained by the pair of PET/CT scans [9,10]. However, the current work focuses on a single PET/CT scan analysis to provide relevant oncological prognostic quantitative indicators.

Therefore, the authors consider that an appropriate information source for assessing the performance of the proposed PGOSIs is a specialist-based visual PET evaluation of each patient's oncological state from a consensus of independent experts in the field. In this work, we present a set of quantitative PGOSIs and test their impact at the clinical level by comparing their performance with the corresponding qualitative evaluation carried out by nuclear medicine specialists. We emphasize on the time-efficiency aspect of PGOSI computations by comparing expert-guided semiautomatic strategies and completely automatic approaches.

Materials and methods

The proposed framework for the performance assessment of PGOSI candidates in breast cancer patients is as follows. A set of 100 WB PET/CT scans corresponding to breast cancer patients with different tumor stages were acquired from the Nuclear Medicine Department at Hospital de Sant Pau (Barcelona, Spain) following all international PET/CT imaging acquisition protocols.

Fig. 1



Cancer stage distribution (low, moderate, high, very high) of the patient samples.

These patients were grouped into four categories according to their tumoral state by the consensus of three independent nuclear medicine physicians as shown in Fig. 1, following visual inspection criteria. As the role of the proposed PGOSIs would have a major clinical impact on the prognosis and management of early-stage cancer patients [11], a larger number of patients in this stage were acquired.

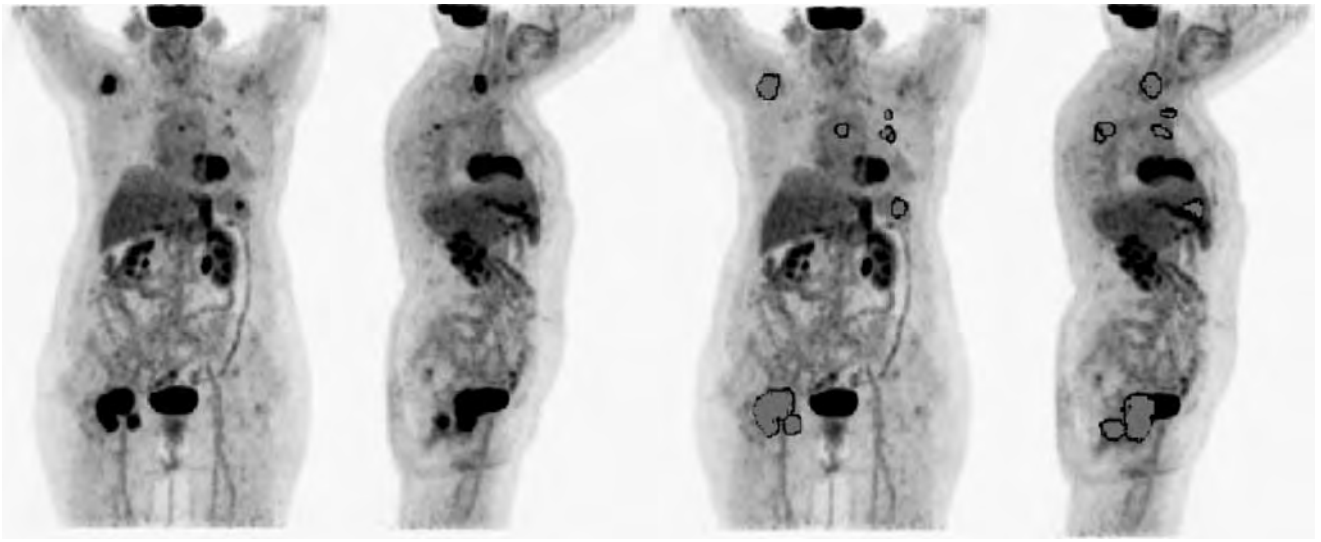
Also, semiautomatic segmentation of metabolic tumor volume obtained from all WB PET scans was carried out by three independent nuclear medicine physicians, which yielded three independent segmentations (S1, S2, and S3) of all patients in order to test the efficacy of any PGOSI proposal in this procedure (Fig. 2). The segmentation was accomplished using a specific-purpose WB PET segmentation software tool.

To be able to test the performance of any given PGOSI proposal, the set of 100 patients was ordered according to their oncological state upon agreement among three independent nuclear medicine experts. The set of clinical variables that were taken into consideration by the experts during the ordering procedure were:

(1) C1: Total tumor volume.
(2) C2: Global aggressiveness of the tumor.
(3) C3: Spread of the tumor – that is, number of organs affected by the tumor and the number of metastases.

Now, given a PGOSI calculation proposal, once computed to the whole set of breast cancer patients, an ordering

Fig. 2



Sample metabolic tumor volume segmentation carried out by a nuclear medicine expert.

of patients according to their PGOSI value can be obtained. The performance (related to clinical impact) of the proposed PGOSI can be addressed by computing the correlation value between the experts' patient ordering and the proposed PGOSI ordering. Figure 3 shows an ordering example of a subset of the breast cancer patients, as well as the corresponding schematic drawings that have been used throughout this paper to illustrate the set of PGOSI proposals and their performance.

The set of PGOSI proposals in this work is detailed as follows. First, it should be noted that, in order to maximize performance, any PGOSI proposal should seek to quantitatively represent the clinical variables that define a patient's tumoral state (C1, C2, and C3). Second, an important property of any PGOSI should be its level of independence from any specialist evaluation, in the sense that an ideal PGOSI should be automatically computed from any given WB PET/CT scan. However, current technology is unable to automatically identify and segment the entire tumoral volume in a given WB PET/CT scan in a reliable manner (although encouraging results have been shown recently in this respect [12]). Thus, in this work we conduct a comparative analysis of the performance of the PGOSI proposals when they are computed in a completely automatic manner or after providing an expert-guided semiautomatic tumor segmentation mask. In doing this, we assume that technological advances will at some point bring both performances to the same level.

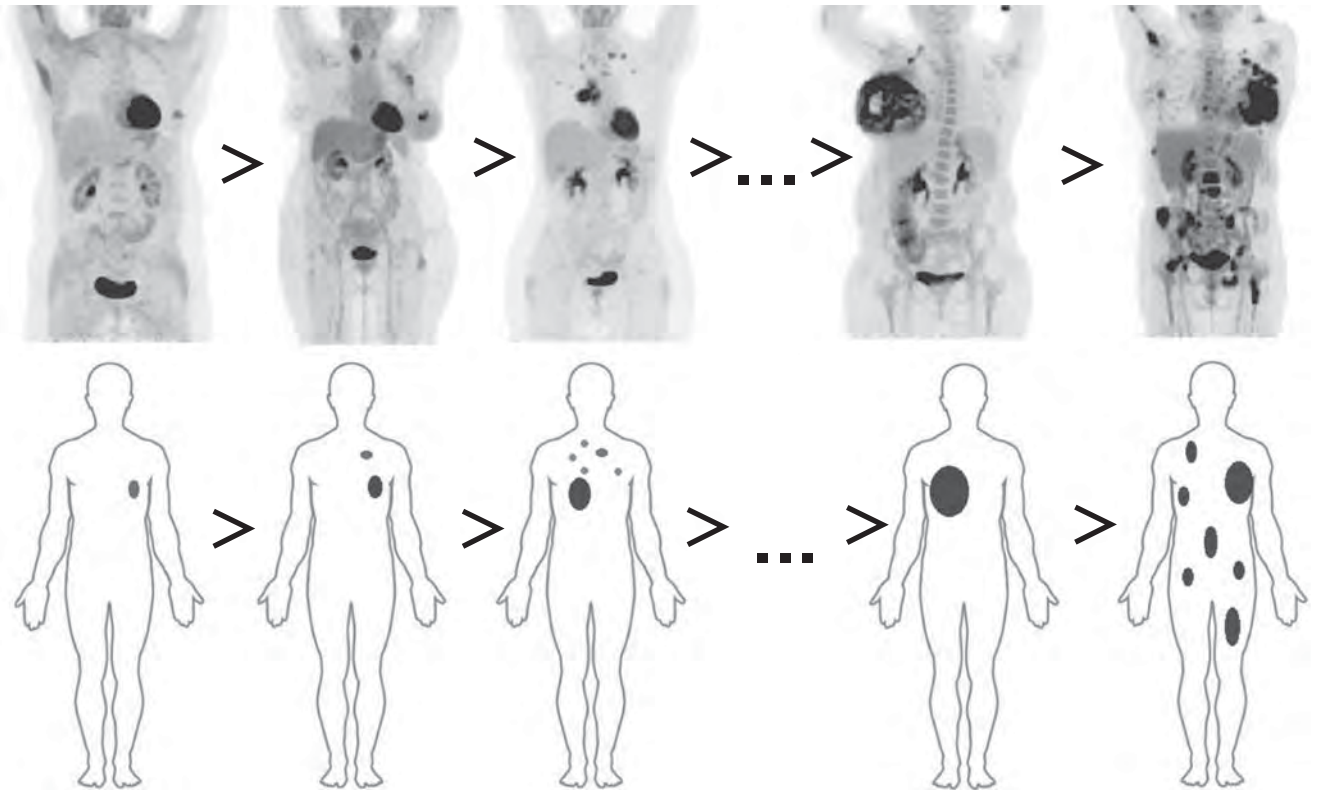
The set of existing clinically justified PGOSI proposals that are related to C1 and C2 have been already mentioned (SUV_{max}, SUV_{mean}, WBMTV, and TLG). The SUV_{max} and SUV_{mean} of the patient's whole-body tumor volume try to measure the cancer aggressiveness but miss the information related to the actual tumor quantity that is present within the patient's body. In contrast, WBMTV does measure the cancer quantity but fails to model its aggressiveness. TLG takes into account both the quantity and aggressiveness of the patient's cancer, but fails to reveal its spread (C3). The limitation of this set of PGOSIs is illustrated in Fig. 4.

To overcome these limitations, a new set of PGOSIs is proposed and described as follows. A key issue to be addressed is how to quantify the cancer spread throughout the patient's body (C3) and compare it with that of other patients, assuming that all of them have the same tumor quantity and the same mean aggressiveness. In particular, a major goal is to be able to distinguish between both tumoral conditions seen in Fig. 4d.

A first alternative would be to compute the number of connected components (NCC) [13] from the PET tumor segmentation mask, which will be related to the number of tumor lesions within the patient's body. This parameter would give a clue about the cancer spread but suffers from some limitations in the special case of a relatively condensed group of tumor lesions, wherein it could be argued that the overall cancer spread would be inferior and therefore the PGOSI value would be so. This limitation is shown in Fig. 5.

To avoid this problem, the NCC value could be combined with the average distance between components, which can be computed by averaging the distance (measured in millimeters, for instance) between the middle points of all connected components. Setting up a new parameter

Fig. 3



Experts' patient ordering in ascending tumoral grade based on clinical visual and semiquantitative variables (top). Schematic illustrations that model the corresponding tumor distribution within each patient's body (bottom) and its aggressiveness (represented by its grayscale intensity: the darker the higher).

based on the product of NCC and average distance between components (aNCC) overcomes the NCC limitation, but produces another limitation related to the average operation, as illustrated in Fig. 6.

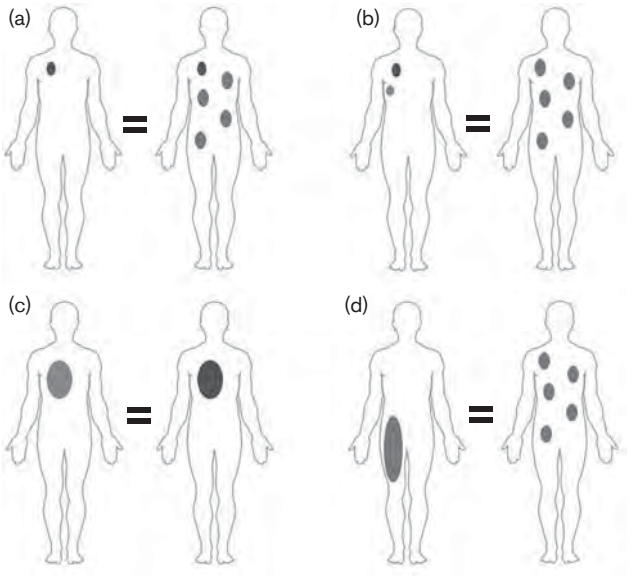
One could try to extend this reasoning by introducing the SD between the middle points of the connected components as a new parameter to be taken into account. However, it rapidly becomes clear that what is needed is a new parameter that approximates the number of organs where the cancer is present within the patient's body. Note that this parameter, referred to as NORG, would overcome the limitations of NCC and aNCC (Fig. 7).

To deal with this task, the following algorithm for obtaining an approximation of the number of affected organs (NORG) is proposed. We start by setting NORG at 0. Then, for each connected component of the PET tumoral segmentation mask, if it has an average HU value significantly different from that of all other connected components or if its middle point is significantly far away from the rest of the connected components, we increase its value by 1. When terminated, an approximation of the actual number of organs or distinct anatomical locations where the tumor is present is obtained. This method has shown a positive correlation of 41% when compared with the NORG value for the 100 breast cancer patients with the actual number of affected organs identified by the medical experts for each patient (where ganglionar adenopathies were considered a single organ except if there existed superior and inferior instances). This result, which is superior to the NCC (31%) or aNCC (33%) correlation, is considered appropriate for quantitatively modeling the patient's cancer spread.

Once a set of several quantitative variables that try to measure the cancer spread has been introduced, to obtain a robust PGOSI proposal, the data obtained from it should be combined with the variables that are related to tumor quantity and aggressiveness.

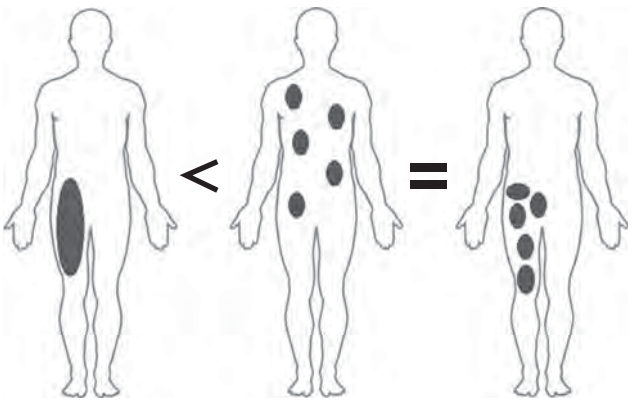
A first step consists of combining the tumor quantity and aggressiveness indicators. TLG has already been proposed for this task, but the authors consider this parameter highly dependent on the segmentation procedure (as it directly includes the WBMTV) and does not consider the distribution of SUVs across all tumor regions (as it includes only the SUV_{mean}).

Fig. 4



Conceptual limitations of common follow-up indicators when used as a PGOSI. The patients' schematic illustrations are based on those defined in Fig. 3. For each case, the equals sign illustrates that the same PGOSI value would be obtained in both patients [(a) SUV_{max}; (b) SUV_{mean}; (c) WBMTV; (d) TLG], albeit possessing an arguably different oncological state. PGOSI, PET Global Oncological State Indicator; SUV, standard uptake value; TLG, total lesion glycolysis; WBMTV, whole-body metabolic tumor volume.

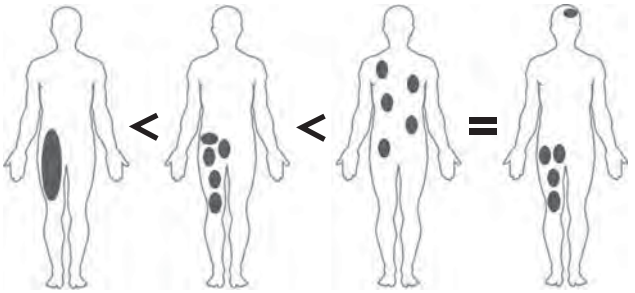
Fig. 5



Limitation of the NCC as a measure of cancer spread. Assuming the same tumoral volume and SUV_{mean}, the NCC parameter is able to distinguish between a single big lesion and a set of smaller lesions, but does not give a clue about their spatial distribution, leading to possible clinical miss-ordering. NCC, number of connected components; SUV, standard uptake value.

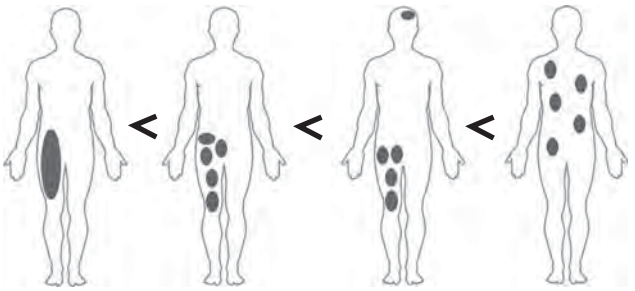
Thus, a new parameter for this measurement is introduced as the sum of all SUVs of the tumor segmentation mask voxels. We consider this number to be less sensitive to the chosen segmentation method as the boundary voxels in tumor lesions (which is generally

Fig. 6



Limitation of the aNCC parameter as a measure of cancer spread. Although it succeeds at distinguishing between common spread differences, the average distance measure could lead to some miss-orderings, as it could be argued that the last tumoral state would be inferior to the third (where the tumor has reached a larger number of distinct anatomical locations).

Fig. 7

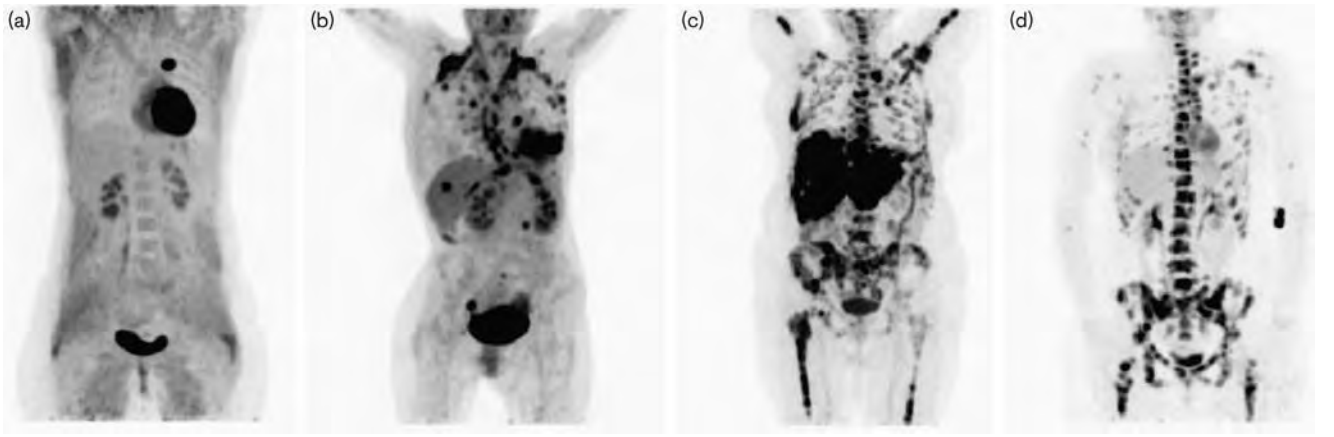


The NORG parameter as a good conceptual indicator of cancer spread across the patient's body.

responsible for the difference among segmentation methods) will contribute less to the final parameter value, as they tend to have a lower SUV. Also, taking the sum and not the average of all tumor SUVs will provide a better sense of distribution of its aggressiveness. This parameter is then normalized by voxel size (in mm^3) and the patient's body surface area, which can be easily obtained from the DICOM scan metadata. We will refer to this described parameter as nTSUV.

Before addressing the performance results of the set of PGOSI proposals, note that its derivation and analysis have been highly simplified, in the sense that in some PET/CT scans its values could be substantially altered by physiological and pathological phenomena. For instance, SUV_{max} and SUV_{mean} may be altered because of muscular uptake (Fig. 8a) or partial volume effects if the lesions are located near brown fat uptake (Fig. 8b) [14]. Also, segmenting false-positive or false-negative ^{18}F-FDG uptakes (e.g. inflammation) could alter most of the indicators, especially the WBMTV value (and even the spread indicators if a false lesion is significantly isolated from the

Fig. 8



A sample of patient's physiological and pathological states that could alter any PGOSI performance. (a) muscular uptake, (b) brown fat uptake, (c and d) advanced tumoral stage. PGOSI, PET Global Oncological State Indicator.

Table 1 Performance results using Pearson's correlation coefficient of a set of PET Global Oncological State Indicator proposals

	Manual segmentation					Auto segmentation	
PGOSI	S1	S2	S3	Mean	SD	MLF	Threshold
SUV_{mean}	0.4929	0.4698	0.4945	0.4857	0.0138	0.3319	0.2492
SUV_{max}	0.5965	0.6142	0.6142	0.6083	0.0102	0.4502	0.2487
WBMTV	0.7997	0.8124	0.8040	0.8054	0.0064	0.5664	0.3154
TLG	0.7934	0.8015	0.7944	0.7964	0.0044	0.5730	0.2733
nTSUV	0.8024	0.8074	0.8026	0.8041	0.0028	0.5759	0.2562
nTSUV*NCC	0.8581	0.8561	0.8528	0.8557	0.0027	0.6214	0.2567
nTSUV*aNCC	0.8429	0.8392	0.8384	0.8402	0.0024	0.6114	0.2563
nTSUV*NORG	0.8712	0.8597	0.8642	**0.8650**	0.0058	**0.6351**	0.2578

Maximum correlation values are highlighted in bold.
MLF, machine learning framework; PGOSI, PET Global Oncological State Indicator; SUV, standard uptake value; TLG, total lesion glycolysis; WBMTV, whole-body metabolic tumor volume.

rest). Finally, the NCC, aNCC, and NORG parameters may not accurately model what they are intended for in advanced tumoral states, as seen in Fig. 8c and d. Since the experts' visual evaluation is not conditioned on these quantitative parameter variabilities, this could reflect a first limitation of the proposed PGOSI framework.

In the next section, an exhaustive performance analysis of a set of PGOSI proposals is presented. All PGOSI proposals were obtained by combining the previously described parameters, which seek to quantify the qualitative information that the medical experts use to evaluate the patients' global oncological state from WB PET/CT scans.

Results and discussion

In this section, performance results of a set of PGOSI proposals in terms of the Pearson correlation coefficient of the experts' ordering of the 100 breast cancer patients and the ordering obtained from the computation of each PGOSI are addressed.

Table 1 shows the correlation results of a set of PGOSI proposals. Performance using either manual (i.e. expert guided) or automatic tumor segmentation techniques is presented. For the manual segmentation scenario, to evaluate the segmentation independence of all PGOSI proposals, performance results are computed in three different tumor segmentation masks segmented by three independent nuclear medicine physicians (S1, S2, and S3). Completely automatic tumor segmentation strategies include the machine learning framework (MLF) described by Sampedro *et al.* [12] and a naïve direct thresholding method at an SUV of 3.0 [15].

First, note the performance results of the state-of-the-art indicators. As predicted in the previous section, the SUV_{mean} (48%) and SUV_{max} (60%) do not model a patient's global tumoral state precisely. The WBMTV and TLG parameters, as expected, give much better results (80%). These results are consistent with those obtained in clinical studies [16–18]. It is to be noted that our proposed nTSUV indicator gives the same correlation performance (80%) but is up to three times more

independent of the manual segmentation used, which is consistent with its design and proposal.

Therefore, the authors consider that the indicator that best models the quantity and aggressiveness of the tumor is the nTSUV. Now, this parameter should be combined with the spread indicators NCC, aNCC, and NORG to improve the performance results. As can be observed, a significant improvement of 6% correlation was achieved. Although no significant difference is shown regarding the use of NCC, aNCC, and NORG, the best performance results were obtained by combining nTSUV and NORG, which is consistent with the derivation presented in the previous section.

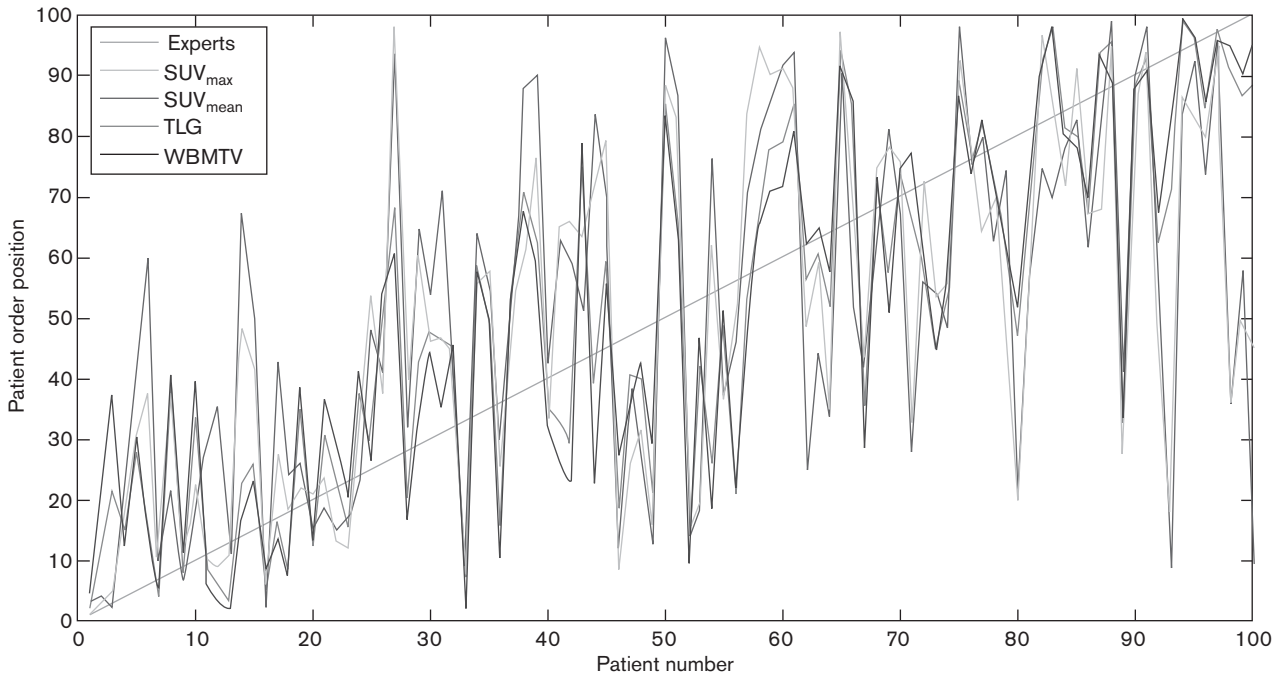
Regarding the results obtained using a completely automatic segmentation scenario, because of the high complexity of the problem, significantly lower correlation results were obtained. Very poor correlation ($< 32\%$) was obtained using the direct thresholding method, which is consistent with the fact that this method would consider any voxel with an SUV greater than 3.0 as tumoral, including physiological uptakes in the brain, heart, kidneys, and bladder. Moderate but significant correlation results were obtained using the MLF method (63%), which were remarkably higher than those of some state-of-the-art indicators such as SUV_{mean} and SUV_{max}. It is noteworthy that this methodology, despite showing about 20% worse performance than the best manual segmentation alternative, is much more time efficient and does not suffer from variabilities due to different segmentations obtained by different experts or software tools.

Figure 9 shows the state-of-the-art visual correlation results between the experts' ordering and each PGOSI using manual segmentation (S1). Figure 10 shows the corresponding results of the nTSUV*NORG PGOSI using either manual segmentation (S1) or automatic segmentation (MLF).

Another way of evaluating the performance of the nTSUV*NORG PGOSI could be by computing its mean number of position errors from the experts' ordering. Using manual segmentation, the value obtained was 11.3 ± 10.2, which means that on average the ordering resolution of this indicator is 11 positions. If one considers a plausible 5% of outliers due to either segmentation or experts' ordering errors, this number reduces to 9.8 ± 7.9. Considering that during the ordering process the nuclear medicine physicians agreed that there would be a mean 5–6-position variance if the ordering was carried out independently instead of by consensus, this result can be considered noteworthy. Using automatic segmentation (MLF), the results were 17.9 ± 17.0 and 15.3 ± 13.1 (if a 5% outlier is assumed).

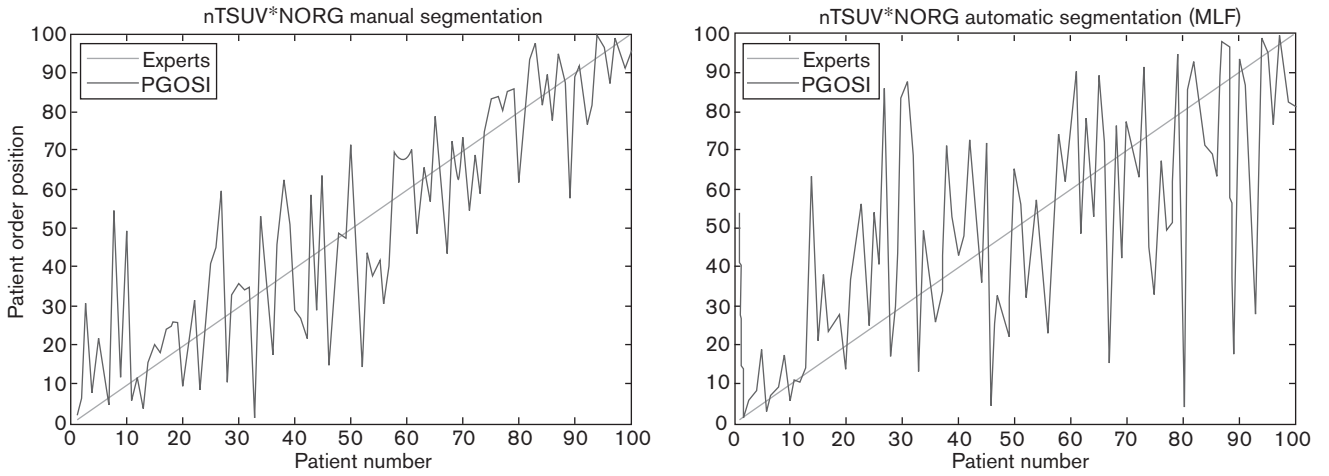
In clinical practice, the impact of the proposed PGOSI can be addressed by establishing a numeric indicator

Fig. 9



State-of-the art PGOSI performance using manual whole-body tumor segmentation. PGOSI, PET Global Oncological State Indicator; SUV, standard uptake value; TLG, total lesion glycolysis; WBMTV, whole-body metabolic tumor volume.

Fig. 10



nTSUV*NORG PGOSI performance results using either manual or automatic whole-body tumoral segmentation. MLF, machine learning framework; PGOSI, PET Global Oncological State Indicator.

Table 2 Performance results using Pearson's correlation coefficient of a set of PET Global Oncological State Indicator proposals

	Manual segmentation					Auto segmentation	
PGOSI	S1	S2	S3	Mean	SD	MLF	Threshold
NCC	0.8251	0.8210	0.8255	0.8239	0.0025	0.6430	0.0012
aNCC	0.7639	0.7891	0.7753	0.7761	0.0126	0.5614	0.0025
NORG	0.8083	0.7994	0.8177	0.8085	0.0092	0.6526	0.0034
WBMTV*NCC	0.8476	0.8471	0.8523	0.8490	0.0029	0.6136	0.3174
WBMTV*aNCC	0.8370	0.8326	0.8346	0.8347	0.0022	0.6094	0.3152
WBMTV*NORG	0.8592	0.8498	0.8571	0.8554	0.0049	0.6300	0.3154
TLG*NCC	0.8519	0.8520	0.8512	0.8517	0.0004	0.6163	0.2713
TLG*aNCC	0.8420	0.8351	0.8340	0.8370	0.0043	0.6091	0.2742
TLG*NORG	0.8661	0.8524	0.8565	0.8583	0.0070	0.6334	0.2733

MLF, machine learning framework; NCC, number of connected components; PGOSI, PET Global Oncological State Indicator.

range for each of the four groups of patients based on oncological state (low, moderate, high, very high). In this case, for 90% of patients in the low group, the nTSUV*NORG value range was 21.52–5157.77; for 40% of patients in the medium group the range was 5249.52–16486.45; for 65% of patients in the high group the range was 17852.39–138386.08; and for 71% of patients in the very high group the range was 210293.92–7882691.02. These results are consistent with the difficulty of distinguishing between medium and high oncological states in a quantitative manner with a relatively small patient sample.

For the sake of completeness, Table 2 shows the performance of another set of PGOSI proposals based on the combination of other relevant indicators that have been described in this work. Although none of them achieved the performance of nTSUV*NORG, very similar performance results and tendencies were seen, which confirms that when WBMTV or TLG is combined with cancer spread indicators, a significant performance improvement is obtained.

Finally, a small illustrative test to validate the potential value of the proposed scoring system was conducted. First, the PET/CT scans of five patients (independent from the ones used in the previous analysis) with a very similar oncological state (i.e. in the same stage) were given to three independent nuclear medicine specialists to be ordered according to oncological state. As expected, the ordering varied among the specialists, and therefore the possible ordering obtained by consensus among all of them may be weak. Then, in the same setting, a new independent set of five patients in a very similar oncological state was selected. Now, however, not only images but also some of the PGOSI values for each patient (in particular, the WBMTV and nTSUV*NORG values) were provided to the specialist. A substantial agreement in the ordering by the three specialists compared with the previous scenario was

observed, which would induce a more robust ordering by consensus.

In summary, the presented results show how quantitative indicators that model the patient's oncological state from a WB PET/CT scan can be obtained such that there is significant agreement with the corresponding human-expert visual analysis. This fact represents an important contribution, as numerical indicators are known to be much more convenient in decision-making scenarios because of their robustness and human independence.

Conclusion

In this work we have presented a number of quantitative indicators computed from WB PET/CT scans that seek to model the global oncological state of a given patient. The design and performance of the proposed indicators have been addressed through a qualitative evaluation of a set of 100 breast cancer patients from a consensus of three nuclear medicine physicians. In this process, the specialists took into consideration visual and semiquantitative parameters related to the patient's tumor volume, aggressiveness, and spread. Therefore, the set of proposed quantitative indicators have been designed to model these tumor properties through the computational image analysis of the metabolic tumor volume segmentation of a WB PET scan, aiming to maximize independence from specialist evaluation.

Performance results based on the correlation between the ordering by global tumoral state of the 100 breast cancer patients performed by the consensus of experts and the proposed quantitative indicators have shown up to 87% correlation using expert-guided PET tumor volume segmentation and 64% using a completely automatic segmentation framework.

The authors consider that the results of this work have contributed to support the need of a quantitative oncological summary of a WB PET/CT scan, which would prove helpful in oncological decision-making scenarios when combined with other cancer indicators. Future work includes performing case studies in different cancer types in which PET evaluation plays a significant role (e.g. lymphoma, sarcoma, or ovarian cancer), as well as keeping track of automatic PET tumor segmentation technologies to obtain a reliable, time-efficient, and expert independent indicator computation system.

Finally, all the described framework and results will need to be validated in large cohorts in long-term studies to fully determine whether the proposed indicators are useful in oncological and nuclear medicine settings to address the prediction of the patient's outcome and treatment response.

Acknowledgements

The work of Frederic Sampedro is supported by the Spanish government FPU (Formación del Profesorado Universitario) doctoral grant. The authors express their appreciation to Professor Ignasi Carrio, Director of the Nuclear Medicine Department of the Hospital de Sant Pau (Barcelona, Spain), for providing the opportunity to develop this work.

Conflicts of interest

There are no conflicts of interest.

References

1 Hwan S, Hyup S, Young J. Prognostic significance of volume-based PET parameters in cancer patients. *Korean J Radiol* 2013; **14**:1–12.
2 Fonti R, Larobina M, Del Vecchio S, De Luca S, Fabbricina R, Catalano L, *et al*. Metabolic tumor volume assessed by ^{18}F-FDG PET/CT for the prediction of outcome in patients with multiple myeloma. *J Nucl Med* 2012; **53**:1829–1835.
3 Chung MK, Jeong HS, Park SG, Jang JY, Son YI, Choi JY, *et al*. Metabolic tumor volume of [^{18}F]-fluorodeoxyglucose positron emission tomography/computed tomography predicts short-term outcome to radiotherapy with or without chemotherapy in pharyngeal cancer. *Clin Cancer Res* 2009; **15**:5861–5868.
4 Hyun SH, Choi JY, Shim YM, Kim K, Lee SJ, Cho YS, *et al*. Prognostic value of metabolic tumor volume measured by ^{18}F-fluorodeoxyglucose positron emission tomography in patients with esophageal carcinoma. *Ann Surg Oncol* 2010; **17**:115–122.
5 Yan H, Wang R, Zhao F, Zhu K, Jiang S, Zhao W, *et al*. Measurement of tumor volume by PET to evaluate prognosis in patients with advanced non-small cell lung cancer treated by non-surgical therapy. *Acta Radiol* 2011; **52**:646–650.
6 Zhu D, Ma T, Niu Z, Zheng J, Han A, Zhao S, *et al*. Prognostic significance of metabolic parameters measured by (18)F-fluorodeoxyglucose positron emission tomography/computed tomography in patients with small cell lung cancer. *Lung Cancer* 2011; **73**:332–337.
7 Liao S, Penney BC, Wroblewski K, Zhang H, Simon CA, Kampalath R, *et al*. Prognostic value of metabolic tumor burden on (18)F-FDG PET in nonsurgical patients with non-small cell lung cancer. *Eur J Nucl Med Mol Imaging* 2012; **39**:27–38.
8 Chan W, Mak H, Huang B, Yeung D, Kwong D, Khong P. Nasopharyngeal carcinoma: relationship between ^{18}F-FDG PET-CT maximum standardized uptake value, metabolic tumour volume and total lesion glycolysis and TNM classification. *Nucl Med Commun* 2010; **31**:206–210.
9 Takeda A, Yokosuka N, Ohashi T, Kunieda E, Fujii H, Aoki Y, *et al*. The maximum standardized uptake value (SUV_{max}) on FDG-PET is a strong predictor of local recurrence for localized non-small-cell lung cancer after stereotactic body radiotherapy (SBRT). *Radiother Oncol* 2011; **101**:291–297.
10 Zhang H, Wroblewski K, Liao S, Kampalath R, Penney B, Zhang Y, Pu Y. Prognostic value of metabolic tumor burden from ^{18}F-FDG PET in surgical patients with non-small-cell lung cancer. *Acad Radiol* 2013; **20**:32–40.
11 Garami Z, Hascsi Z, Varga J, Dinya T, Tanyi M, Garai I, *et al*. The value of 18-FDG PET/CT in early-stage breast cancer compared to traditional diagnostic modalities with an emphasis on changes in disease stage designation and treatment plan. *Eur J Surg Oncol* 2012; **38**:31–37.
12 Sampedro F, Escalera S, Domenech A, Carrió I. Automatic metabolic tumor volume segmentation in breast cancer whole-body PET/CT scans: a supervised learning approach. *Med Image Anal* 2013. Under revision.
13 Biancardi A, Segovia M. Adaptive segmentation of MR axial brain images using connected components. *Lect Notes Comput Sci* 2001; **2059**:295–302.
14 Zhang H, Wroblewski K, Appelbaum D, Pu Y. Independent prognostic value of whole body metabolic tumor burden from FDG-PET in non-small cell lung cancer. *Int J Comput Assist Radiol Surg* 2013; **8**:181–191.
15 Oh J, Seo J, Chong A, Min J, Song H, Kim Y, Bom H. Whole-body metabolic tumour volume of ^{18}F-FDG PET/CT improves the prediction of

prognosis in small cell lung cancer. *Eur J Nucl Med Mol Imaging* 2012; **39**:925–935.
16 Lee P, Weerasuriya DK, Lavori PW, Quon A, Hara W, Maxim PG, *et al*. Metabolic tumor burden predicts for disease progression and death in lung cancer. *Int J Radiat Oncol Biol Phys* 2007; **69**:328–333.
17 Larson SM, Erdi Y, Akhurst T, Mazumdar M, Macapinlac HA, Finn RD, *et al*. Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging. The visual response score and the change in total lesions glycolysis. *Clin Positron Imaging* 1999; **2**:159–171.
18 Borst GR, Belderbos JS, Boellaard R, Comans EF, De Jaeger K, Lammertsma AA, Lebesque JV. Standardised FDG uptake: a prognostic factor for inoperable non-small cell lung cancer. *Eur J Cancer* 2005; **41**:1533–1541.

Computers in Biology and Medicine 55 (2014) 92–99

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

A computational framework for cancer response assessment based on oncological PET-CT scans

Frederic Sampedro [a,*], Sergio Escalera [b], Anna Domenech [c], Ignasi Carrio [c]

[a] Autonomous University of Barcelona, Faculty of Medicine, 08193 Barcelona, Spain
[b] Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Barcelona, Spain. Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via de les Corts 585, 08007, Barcelona, Spain
[c] Hospital de Sant Pau, Nuclear Medicine Department, 89 Carrer Sant Quintí, 08026 Barcelona, Spain

ARTICLE INFO

Article history:
Received 3 June 2014
Accepted 13 October 2014

Keywords:
Computer aided diagnosis
Nuclear medicine
Machine learning
Image processing
Quantitative analysis

ABSTRACT

In this work we present a comprehensive computational framework to help in the clinical assessment of cancer response from a pair of time consecutive oncological PET-CT scans. In this scenario, the design and implementation of a supervised machine learning system to predict and quantify cancer progression or response conditions by introducing a novel feature set that models the underlying clinical context is described. Performance results in 100 clinical cases (corresponding to 200 whole body PET-CT scans) in comparing expert-based visual analysis and classifier decision making show up to 70% accuracy within a completely automatic pipeline and 90% accuracy when providing the system with expert-guided PET tumor segmentation masks.

1. Introduction

^{18}F-fluorodeoxyglucose (^{18}F-FDG) positron emission tomography (PET) has become a standard imaging method for the staging, restaging, and monitoring of treatment response in a variety of tumors. By injecting the ^{18}F-FDG (fluorodeoxyglucose) radiopharmaceutical to the patient, a metabolic activity volume, measured in SUV (standard uptake value [1]) units, is acquired. FDG-avid tumors such as lymphoma, sarcoma, breast cancer or ovarian cancer show higher than normal SUV values in non-physiological locations, leading to a more accurate diagnosis than MR (magnetic resonance) imaging or CT (computed tomography) in some oncological scenarios [2]. Current technology offers integrated PET-CT and more recently PET-MRI scanners, which provide co-registered PET and CT/MR scans of the patient [3].

This technology has proven especially useful in patient's global cancer response assessment [4,5], where a comparative analysis between two time consecutive whole body PET-CT scans can provide an accurate insight of the morphological and physiologic cancer evolution trends. Generally, nuclear medicine physicians assess a patient's cancer progression or response (Fig. 1) condition following a trained visual analysis of both scans.

By segmenting the tumor volume from the PET scans, changes in metabolic tumor volume (MTV) and its metabolic activity (typically modeled by its mean or maximum SUV values) have shown to be a valuable quantitative indicator of the cancer evolution stage, as described in Refs. [6,7]. However, these studies focus on a particular subset of cancer scenarios, where typically only one significant tumor lesion is analyzed, and changes in the tumor spread over time are not taken into consideration. Sampedro et al. recently showed in [8] that the cancer spread information should be taken into consideration in the quantitative analysis of the oncological state from PET scans.

Note that as obtaining an accurate expert-guided tumor segmentation of whole body PET scans is highly time-inefficient in the clinical day-to-day setting, sometimes a very rough approximation of the volume and activity of each tumor lesion is obtained by placing a user-variable radius sphere on top of the lesion and reading its diameter (commonly in mm) and the mean and maximum metabolic activity values (SUV_{mean}, SUV_{max}) within the sphere's volume.

In this work we introduce a computational framework for the analysis of the cancer time evolution based on two time consecutive PET-CT scans. The aim of this system is to aid in the decision making process regarding the cancer progression or response condition by providing supporting quantitative information from image analysis and machine learning techniques.

Despite being a highly challenging computational scenario (as shown in subsequent sections), nuclear medicine experts agree on the need of such a computational system that could provide objective and quantitative information to support the visual analysis, which is well-known to suffer from inter- and intra-observer variabilities [9].

* Corresponding author.
E-mail addresses: fredsampedro@gmail.com (F. Sampedro), sergio@maia.ub.es (S. Escalera), adomenech@santpau.cat (A. Domenech).

http://dx.doi.org/10.1016/j.compbiomed.2014.10.014
0010-4825/

To the best of our knowledge, the computational modeling of the described scenario has not been addressed by the scientific community. Although nuclear medicine software stations include several tools to carry out expert-guided PET segmentation and allow the direct superimposition of segmentation masks from one PET scan to another, they lack a comprehensive computational decision making and quantitative framework analogous to the one proposed in this work.

The rest of the paper is organized as follows. Section 2 describes the materials used for the validation of the proposed system. Section 3 presents an in-depth characterization of the methodology proposed for the implementation of the computational framework. Section 4 shows the results of the proposed system (using semi-automatic and automatic configurations) and discusses its possible applications in the clinical setting. Finally, Section 5 concludes the paper and points out some future work required to fully validate the proposed system in the clinical domain.

2. Materials

A total of 200 whole body FDG PET-CT scans were obtained from the Philips PET-CT Gemini TF machine located at the nuclear medicine department in the Hospital de Sant Pau (Barcelona, Spain). Each scan contains two co-registered volumes (PET and CT) in DICOM (Digital Imaging and Communications in Medicine) format. From the DICOM metadata, SUV values for PET voxels and HU (Hounsfield units [10]) values for CT voxels can be computed. A PET voxel corresponds to 64 mm^3 and a CT voxel to 2 mm^3. Volume dimensions are $144 \times 144 \times Np$ for PET volumes and $512 \times 512 \times Nc$ for CT volumes (Np and Nc being the number of slices for each volume). Np and Nc (ranging from 192 to 213 and 511 to 623, respectively) are related with a ratio of $Nc/Np = 2.66$. However, the actual number of slices is dependent on the volume of interest selected by the acquisition technician, varying with the patient's height and the anatomical limits of interest (typically either from neck to middle-thigh or from the top of the skull to the feet). The patient's position during acquisition is also variable (mainly related to arm positioning).

These 200 scans correspond to 100 patients with Non-Hodgkin lymphoma or breast cancer (where the proposed framework would be of much clinical interest), each one having two time consecutive scans (T and $T+1$) with which oncological evolution is addressed. The time elapsed between scans is typically between 3 and 8 months, depending on external clinical factors such as the type of treatment received and other clinical and logistic variables.

The ground truth information regarding clinical condition for each case was given following a consensus of three independent nuclear medicine physicians (53 cancer progression and 47 cancer response conditions). The same consensus also agreed conceptually about the existence and location of tumor lesions in all PET scans. Subsequently, each physician segmented a random third of the scans, therefore providing the expert-guided tumoral segmentation masks of 200 PET scans. Note that in an analogous context, Sampedro et al. [11] showed good inter-observer segmentation overlap degree in segmenting oncological PET scans, and therefore this issue is not considered in this work.

The disease extent was variable across subjects, ranging from a single small tumor lesion to a highly spread metastatic tumor. However, the majority of the cases (approximately 80%) showed multiple tumor lesions (analogous to the scans in Figs. 1a and 2), where the decision making process becomes more difficult in clinical

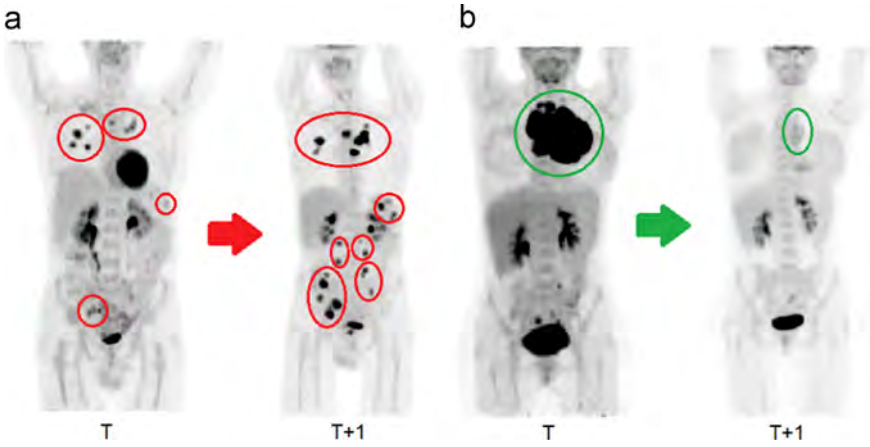


Fig. 1. Sample cancer progression (a) and response (b) conditions as shown by two time consecutives PET scans. Maximum intensity projections (MIP) of the PET volumes are displayed.

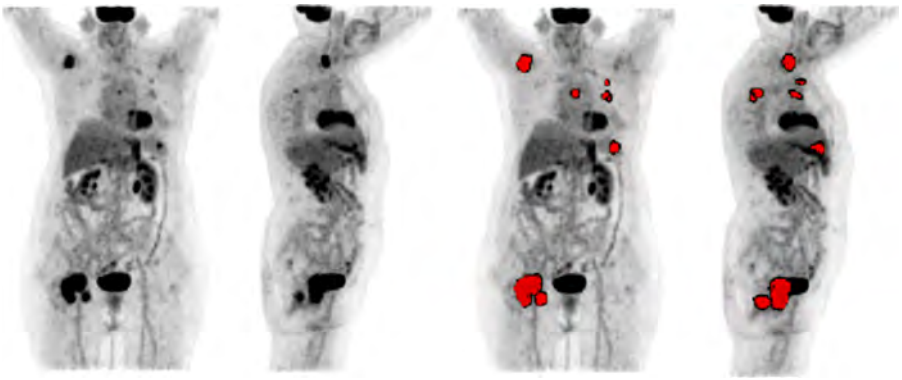


Fig. 2. Whole body FDG-PET MIP projections of a patient (left) and its corresponding tumor volume segmentation carried out by a nuclear medicine physician (right).

practice and automatic segmentation techniques would be valuable, as the semi-automatic approach would be highly time-consuming.

3. Methods

In this section we first present the proposed computation model of the stated clinical scenario at a system level and then describe its implementation based on a supervised machine learning framework.

3.1. Computational model and system design

In general, there exist several clinical scenarios that may be present when analyzing the cancer evolution condition from a particular pair of time consecutive PET scans, which base cases are summarized in Fig. 3.

The most common cases are also the most intuitive, which are shown in Fig. 3a–c. As an example, the cancer progression shown in Fig. 1a can be modeled as a combination of the progression conditions in Fig. 3a, b and d, as in time $T+1$ the tumor lesions that were present at time T increased in size and metabolic activity and new lesions appeared throughout the subject's body. Analogously, the cancer response condition shown in Fig. 1b can be associated with a combination of Fig. 3a and b conditions, as the tumor shrank in size and decreased in SUV value. Finally, examples of Fig. 3d and e conditions are shown in Fig. 4.

Once the clinical cases that are used to identify a cancer progression or response condition have been modeled, the clinical problem can be simplified to a pattern recognition problem where a binary decision has to be made, providing as input two time-consecutive PET scans from the same patient. Fig. 5 shows the block diagram of the proposed fully automatic computational system to address this problem.

Two modules are required to successfully identify the cancer evolution condition from a pair of PET-CT scans. First, the tumor segmentation masks need to be obtained by an Automatic Tumor Segmentation module which, as will be described in Section 3.2 represents a challenging computational problem and therefore can be identified as a performance bottleneck of the problem. Second, an automatic decision making module is responsible for recognizing the clinical patterns illustrated in Fig. 3 (or a combination of thereof) and output the predicted cancer evolution condition. Note that the incorporation of an additional module to quantify the "intensity or severity" (in terms yet to be defined) of a particular

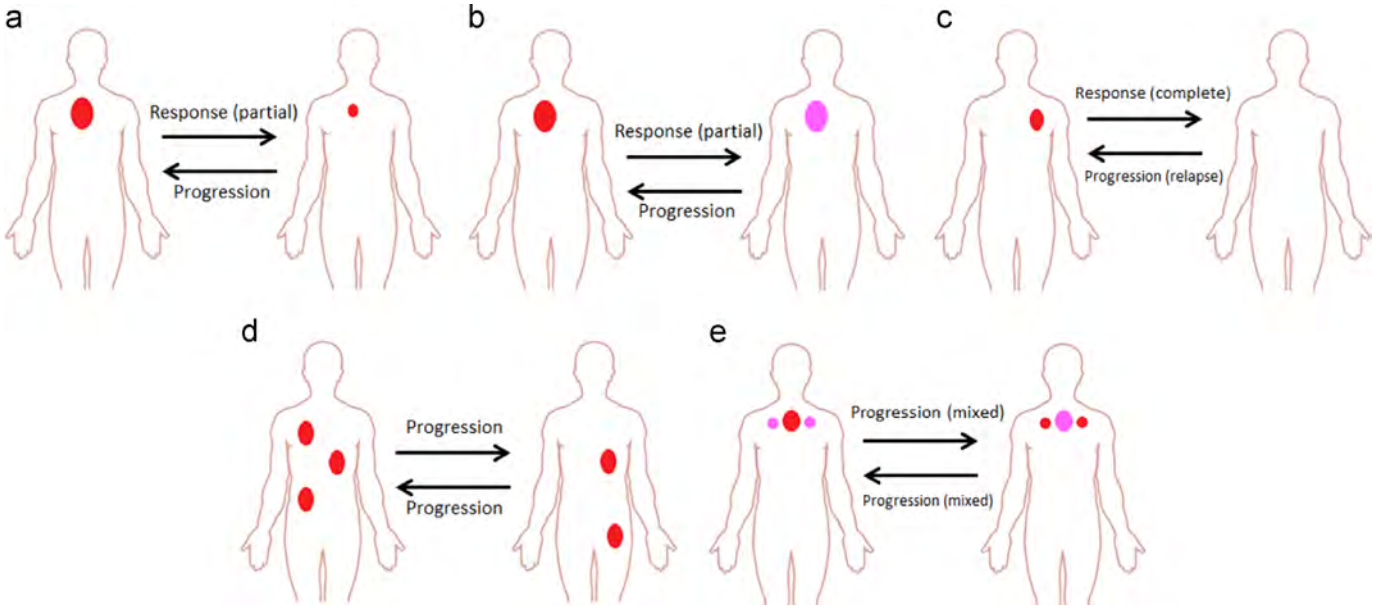


Fig. 3. Illustrations of the typical cancer evolution scenarios in nuclear medicine. When a single tumor lesion is present, its change in volume or intensity in time define its progression or response condition (a, b, c). In the multi-lesion case, the spread of the cancer into new anatomical locations (regardless of the volume change) or the intensity increase in any of the lesions is clinically associated with a progression scenario (d, e). Intensity of the tumor lesion is represented by its color intensity (pink-low, red-high). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 4. Real examples of the cancer evolution conditions illustrated in Fig. 3d (a) Fig. 3e (b). Note the appearance in (a) of new tumoral lesions in time $T+1$ with respect to T and how in(b) most lesions decrease in intensity but the one pointed by the open arrow shows an increase.

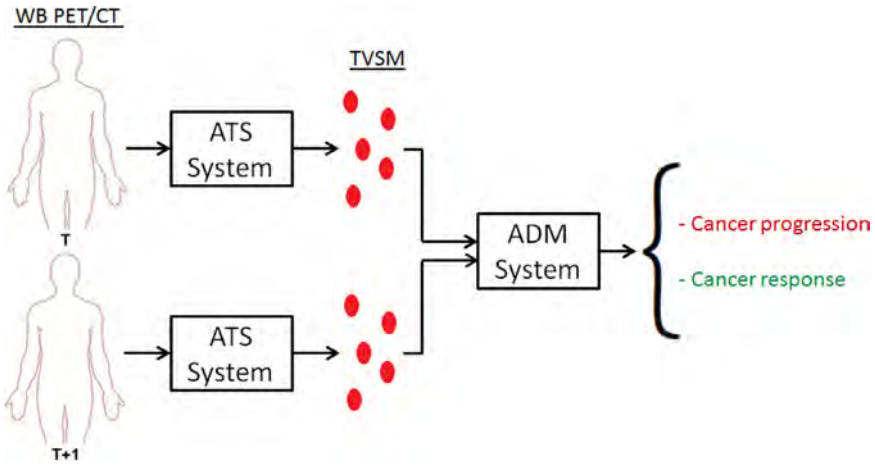


Fig. 5. Computational framework block diagram. WB PET-CT: whole body PET-CT scan, ATS: automatic tumor segmentation, TVSM: tumor volume segmentation Mask, ADM: automatic decision making.

cancer progression or response would raise the potential clinical value of the system. However, the sole computational ability to accurately predict the type of cancer evolution represents a challenging enough problem to be addressed in this work.

Given the high complexity of the problem under consideration, the proposed system, albeit proving time-efficient and objective information, may produce a significant amount of incorrect results at any level (both within the ATS and ADM modules). Therefore, a semi-automatic approach where an expert may validate the process carried out by the computational system at any level may be a reasonable setting, as the accuracy of the overall framework would increase and the expert may receive clinically relevant quantitative information from the system modules.

3.2. *System implementation: A supervised machine learning framework*

In this subsection the implementation of the system blocks in Fig. 5 is presented. A key point to note is the high complexity of the whole system. On one hand, the ATS system should be able to discriminate between tumoral and physiological volumes within any whole body PET scan, which requires the correct identification of the heart, bladder, kidneys and brain among others, which can show substantial differences in morphology and intensity patterns between a healthy infant and an eldest person with an advanced oncological state. On the other hand, the ADM system can be faced with a huge variety of cancer evolution patterns, from which it is responsible for the identification of some of the conditions illustrated in Fig. 3. Furthermore, there is no general agreement on the magnitude of an SUV change to be considered as significant to model a progression or a response (Fig. 3b and c), that is, a small change in SUV_{mean} between 1% and 10% may not be considered enough for discrimination. As the whole decision making process carried out by the medical experts relies on a combination of mainly visual analysis features, the implementation of a direct and analytical mathematical model would be too restrictive.

These properties of the computational setting lead to the adoption of a supervised machine learning based solution for both system blocks. An ATS system was already designed and implemented in Sampedro et al. [11] where the authors deal with all the specific challenges of the problem by building an ad-hoc PET voxel feature set that given enough training data can recognize most pathological and physiological patterns (from specific organ normal uptake to other phenomena such as muscular uptake, brown fat uptake or inflammation). In short, in [11] the authors proposed an ATS system derived from modeling a set of clinical facts that are related to the presence of tumor tissue in a computational PET voxel feature set and obtained an expert-guided training set of PET volumes which was used to train a supervised learning classifier capable of providing a tumor segmentation mask proposal from any given whole body PET scan. Note that this alterative offers conceptual advantages to the few other existing fully automatic whole body PET tumor segmentation techniques, which assume tumor homogeneity and "hot spot" structure [19]. In this work we will use this system as a black box module for the automatic extraction of the PET tumoral segmentation masks.

Therefore, the remaining portion of this subsection will present in depth the design and implementation of our ADM system. First, given that it will be based on a machine learning scheme, an ad-hoc feature set must be obtained from the ADM input data (both T and $T+1$ PET tumor segmentation masks). The main design goal of this feature set is to maximize its capacity to model appropriately the underlying scenario and therefore help the supervised learning algorithms to accurately detect the desired patterns and correctly discriminate between a cancer progression and response condition.

The chosen feature set is described next. On one hand, consider a base case with the presence of a single tumor lesion in both T and $T+1$ scans (Fig. 3a–c). In this scenario, changes in total tumor volume (ΔV_T) and its SUV_{mean} (ΔSUV_{mean}) are probably the most relevant descriptors of the underlying phenomena. We have also considered the incorporation of the change in SUV_{max} (ΔSUV_{max}) as an appropriate descriptor, as it has proven to be a relevant indicator in the clinical literature [12] and at the computational level could account for the alteration of the ΔSUV_{mean} indicator due to segmentation variations between both masks.

On the other hand, it is clear that additional descriptors are needed to model the cases illustrated in Fig. 3d and e. To do so, the multi-lesion scenario must be addressed. First, in order to do it within an image processing framework, the number of connected components (NCC) in the segmentation mask is used to model the number of tumor lesions present within the patient's body at any time as described in [8]. Then, the change in NCC between time T and $T+1$ masks (ΔNCC) is introduced as a descriptor in the feature set. A particular limitation is that although this is probably the most sensible approach, the NCC parameter can be "noisy", meaning that a particular single lesion may be segmented in different connected components depending on the segmentation method employed (especially if using threshold-based methods).

Note that although the ΔNCC parameter provides useful information for the decision making process regarding the global change of tumor lesions in time, it is still not enough to model the evolution cases illustrated in Fig. 3d and e (i.e., a negative or nil ΔNCC can still reflect a progression condition). To overcome this final limitation, two

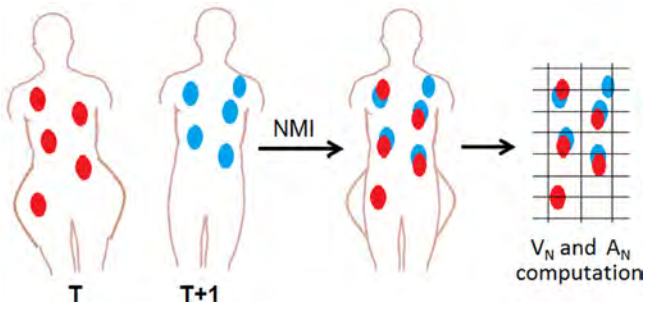


Fig. 6. Illustration of the coregistering process needed to spatially compare the T and $T+1$ segmentation masks in a robust manner. NMI: normalized mutual information.

additional features are added to the final descriptor. The first is responsible of modeling the appearance of new tumoral lesions in time $T+1$ with respect to time T, that is the detection of the tumor volume that spread to a different anatomical location in time (V_N). The computation of V_N is not direct. First, in order to spatially compare both segmentation masks in a robust manner, they must be coregistered in space, since the patient's position and physical shape (severe loss weight is a common cancer symptom) may have considerably changed in both PET scans. In particular, we use the Normalized Mutual Information (NMI) technique to perform this operation (Fig. 6).

Second, both coregistered masks are smoothed (and rebinarized) in order to compensate for small segmentation artifacts and to be able to robustly compare the voxel wise overlaps between specific connected components from both masks. Let the resulting masks be S_T and S_{T+1}. Since they are both binary (logical) 3D matrices, the following element-wise operation can be computed:

$$M_N = S_{T+1} \cap \neg S_T$$

Now, the M_N can be used to detect the set of new tumoral lesions and compute its volume. However, a last processing step should be carried out in order to avoid taking into account the positive voxels in M_N that are "semantically wrong", meaning that they are actually associated to a shared tumor lesion but appear in M_N due to segmentation differences, slight changes in lesion morphology across time or coregistering effects (Fig. 6 right). Therefore, they need to be excluded in the computation of V_N. To accomplish this, Algorithm 1 is applied.

$V_N = 0$

For each non-zero voxel v in M_N:

Compute the connected component of the S_{T+1} mask where v belongs. Let c_{T+1} be its corresponding binary mask.

Compute the connected component of the S_T mask that has the closest voxel (in (Euclidean distance) to v. Let c_T be its corresponding binary mask.

Compute the Jaccard overlap index between both connected components:

$$J = \frac{C_{T+1} \cap C_T}{C_{T+1} \cup C_T}$$

If $J < D$ (a predefined threshold):

$V_N = V_N + 1$

Algorithm I. Computation of V_N.

The last feature to be included is intended to model the most subtle clinical evolution scenarios, illustrated in Fig. 3e and b. For that purpose, local increasing of tumor intensity at the lesion level in time $T+1$ with respect to time T are detected and quantified in the descriptor A_N. To accomplish that, we consider the intersection mask:

$$I = S_{T+1} \cap S_T$$

Then, consider the masked PET volumes by I, P_T and P_{T+1}, in SUV voxel units. We define the logical volume:

$$H = P_{T+1} > \alpha P_T$$

which accounts for the tumor voxels that shown a significantly higher (controlled by α) intensity in time $T+1$ with respect to T. Finally, A_N is defined as the number of non-zero entries in H.

To sum up, the feature set of the supervised learning-based ADM system is shown below:

$$F = \{\Delta V_T,\ \Delta SUV_{mean},\ \Delta SUV_{max},\ \Delta NCC,\ V_N,\ A_N\}$$

Note that if the system is used in semi-automatic mode, meaning that the medical expert are assessing and correcting the computations performed by the system modules, the presentation of the particular feature vector as well as all the intermediate volume masks (especially M_N and H), can provide relevant quantitative information to complement the expert's visual analysis process.

Once our proposed feature set has been defined, the choice of the supervised learning algorithm of the ADM system is addressed. First, note that this particular learning problem is challenging, not only due to the high variability of the input and the subtle evolution scenarios to be detected, but also because of the small and subjective training set available (since PET scans are costly and the ground truth information is derived mainly from an expert's visual analysis).

Therefore, a range of learning algorithms is applied to seek for the best supervised learning strategy within this scenario (following the no free lunch theorem [13]). The set of chosen learning algorithms is the following: Naïve Bayes classifier (given the high independence of the individual features), kNN, decision trees (to try to capture a clinically relevant decision rule) [16], neural networks [17], logistic regression (to obtain an output probability instead of binary decision) [14], and state of the art radial basis function SVM [15] and Discrete AdaBoost [14] approaches.

All the required hyperparameters (k from kNN, SVM RBF parameters, number of units/layers of the neural network, number of decision stumps for Adaboost, etc) where chosen using a nested leave-one-out cross validation scheme within the training data, and the external leave-one-out was used at the test phase.

4. Results and discussion

In this section we present the performance results of the proposed system and discuss its implications both at the clinical and technological levels. Two main system usage modes are distinguished, one where the system runs in a completely automatic manner having only as input the pair of time consecutive PET-CT scans of a given patient, and another where the PET tumor segmentation masks carried out by medical experts are also provided as input to the system.

In the first case, the ATS module plays a major role in the overall system performance. As it was mentioned, in this work we use the ATS system trained on 200 independent PET-CT scans described in [11] as a black box module to automatically segment the tumor volume from a give PET-CT scan. For each of the 100 cancer evolution cases described in Section 2, physicians provided the expert-guided PET tumor segmentation masks, allowing us to evaluate the ATS module segmentation performance.

As expected, the segmentation accuracy results at the voxel level are quite low: mean Jaccard overlap index of 0.18 ± 0.21, mean sensitivity (ratio of correctly classified tumor voxels and the sum of correctly classified tumor voxels and incorrectly classified tumor voxels) of 0.23 ± 0.28 and mean specificity (ratio of correctly classified non-tumor voxels and the sum of correctly classified non-tumor voxels and incorrectly classified non-tumor voxels) of 0.9998 ± 0.0005, with a clear tendency to prioritize false negative over false positive decisions (which was intended by the authors as is sensible given the clinical context). However, a more relevant performance indicator of the ATS module is its relative (rather than absolute) discrimination power between the quantities of tumor present in different PET scans. In this respect, our ATS module shows a 71% Pearson correlation coefficient when comparing the total tumor volume of the expert-guided and automatic PET segmentation masks, which is clearly superior to other completely automatic segmentation strategies such as direct thresholding [8] (49%).

Once the system has access to both time T and $T+1$ PET tumor segmentation masks (either from the ATS module or provided by the medical experts), the ADM module executes. The dataset available to address its performance is built from the computation of the six features described in Section 3.2 for each of the 100 cancer evolution cases (where D and α where empirically set to 0.05 and 1.2, respectively). This 100×6 matrix, in conjunction with the ground truth cancer evolution condition (progression or response) for each case is then provided to the set of supervised learning algorithms within a nested leave-one-out cross-validation scheme described in Section 3.2. The tendency of the mean values of the learning algorithm parameters across the cross-validation phase were $k=3$ in kNN, 20 weak classifiers in AdaBoost, $C=1$ (box constraint) and $\sigma=0.05$ (scaling factor in the radial basis function kernel) in SVM, and three layers in NN.

Performance results are shown in Table 1. In this context, accuracy is defined by the fraction of correctly classified evolution cases, sensitivity as the ratio of correctly classified progression cases and the sum of correctly classified progression cases and incorrectly classified response cases, and specificity as the ratio of correctly classified response cases and the sum of correctly classified response cases and incorrectly classified progression cases.

As it can be observed, a substantial performance gap is observed between the automatic and semi-automatic approach, which is associated with the limitations of the ATS module to provide a correct tumor segmentation mask. As an example, consider the cancer progression shown in Fig. 7.

Clearly, the set of features computed from the automatic segmentation masks will be altered due to the errors made by the ATS block and, although these errors may have some patterns (typical errors include the wrong detection of the heart, bladder or kidneys), the learning algorithms may not be able to resolve them when obtaining the classification rule.

Finally, in order to address the convenience of the proposed ADM feature set, the relevance of each feature within the decision making process in the two learning algorithms that achieved the best performance results is shown in Fig. 8. Note that both results are perfectly consistent with the underlying clinical scenario, where the total volume change (ΔV_T) is clearly the most relevant visual indicator and the change in the number of lesions (modeled by ΔNCC) is only relevant if some of the lesions are actually new in time $T+1$ (which is already modeled by V_N).

These results contribute to both computed aided diagnosis and quantitative longitudinal analysis in the medical imaging field. The derivation of quantitative indicators from segmentation procedures to monitor clinically relevant information is common in a number of medical imaging modalities (see [20–22] for examples in MRI, CT, and

Table 1
Accuracy (Acc), sensitivity (SN) and specificity (SP) results of the ADM system at predicting the cancer progression or response condition. Manual (M) or Auto (A) refer to the use of the expert-guided or ATS segmentation masks, and train (Tr) or test (Te) refer to the accuracy obtained in the training or test phases. NB: Naïve Bayes, NNet: neural networks, LReg: logistic regression, DT: decision tree, SVM: RBF support vector machine, kNN: k-nearest neighbors, AB: discrete AdaBoost.

	NB			NNet			LReg			DT			SVM			kNN			AB		
	Acc	SN	SP	Acc	SN	SP	Acc	SN	SP	Acc	SN	SP	Acc	SN	SP	Acc	SN	SP	Acc	SN	SP
MTe	0.77	0.68	0.89	0.8	0.81	0.79	0.83	0.86	0.78	0.85	0.83	0.87	0.85	0.88	0.80	0.86	0.83	0.90	0.90	0.89	0.91
MTr	0.78	0.69	0.90	0.87	0.89	0.85	0.85	0.88	0.80	0.96	0.95	0.97	0.85	0.88	0.80	0.87	0.84	0.89	1.00	1.00	1.00
ATe	0.57	0.34	0.90	0.53	0.54	0.51	0.68	0.77	0.53	0.70	0.69	0.71	0.70	0.71	0.68	0.62	0.64	0.58	0.54	0.51	0.58
Atr	0.58	0.35	0.90	0.70	0.72	0.69	0.71	0.79	0.58	0.88	0.88	0.88	0.74	0.74	0.74	0.76	0.76	0.76	1.00	1.00	1.00

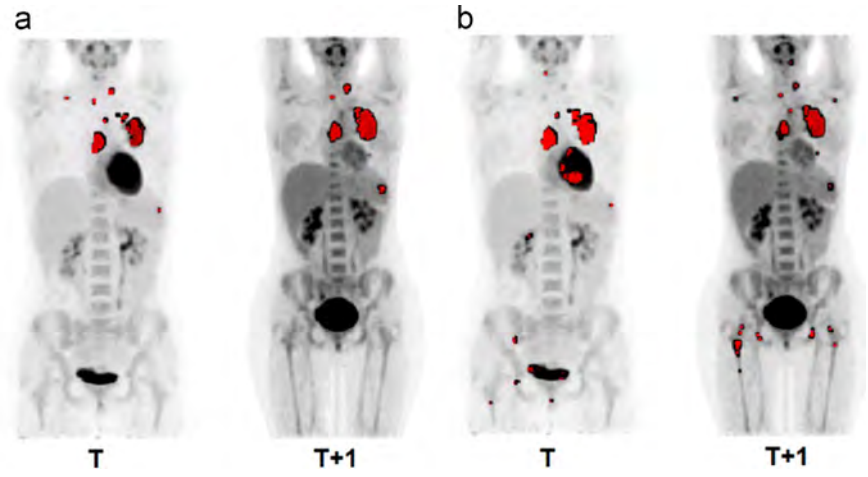


Fig. 7. Illustrative example of the limitations of the ATS system and its possible effects in the ADM performance results. The same cancer progression is shown with the expert-guided tumoral segmentation masks (a) and its corresponding ATS alternatives (b).

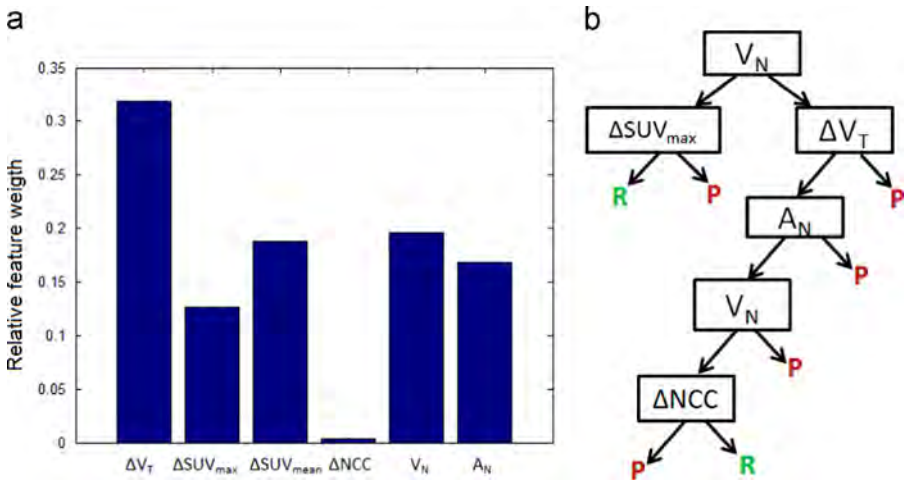


Fig. 8. (a) Feature importance as described by the AdaBoost feature weights. The weight values shown in the graphic corresponds to the accumulated alpha weight for the Adaboost classifier across iterations. (b) Illustration of the decision tree produced by the DT algorithm.

cerebral FDG-PET). Computer aided diagnosis systems and the use of machine learning techniques in them are also emerging within the field [23,24]. In this work, by introducing the proposed framework, we extended the range of applications of these computational techniques in the assessment of cancer evolution in oncological PET-CT scans.

5. Conclusions and future work

We described a common clinical diagnostic scenario in nuclear medicine imaging, the cancer evolution assessment (its progression or response stage) via a pair of time consecutive PET-CT scans, and proposed a computational framework to complement the expert's visual analysis with relevant quantitative and subject-independent information within this diagnostic context.

Since modeling this particular clinical scenario following the high level expert's knowledge and reasoning from a computational point of view represents a challenging problem, a supervised machine learning framework has been proposed. A trade-off is observed between the use of a completely automatic (and therefore time-efficient and subject independent) approach with a performance accuracy of 70% at detecting the correct cancer evolution condition (taking as a reference the expert's visual analysis), and a semi-automatic approach where the system is provided with the PET tumoral segmentation masks carried out by the trained physicians, which show up to 90% performance. In any case, the set of numerical indicators used as feature set within the machine learning framework or the numerical outputs of the trained classifiers may provide the expert with relevant quantitative information, contributing to improve the overall diagnostic accuracy in the clinical setting.

Future work includes, at the clinical level, the incorporation and long term validation of the proposed system in the day-to-day medical practice as well as the introduction of a new quantification module to model the overall intensity of a particular cancer progression or response condition, which would become a very useful tool to help in oncological treatment response analysis and management. At the technological level, we plan on reducing the computation time of the system using GPU-based parallelization techniques as well as dealing with the multi-class problem of recognizing each of the evolution patterns instead of the general response/progression state by introducing successful multi-class frameworks such as error-correcting-output-codes [18].

Conflict of interest statement

None declared.

Acknowledgments

The work of Frederic Sampedro is supported by the Spanish government FPU (Formación del Profesorado Universitario) doctoral grant (Grant No. AP2012-0400). The project has been partially supported by Spanish government project TIN2013-43478-P.

References

[1] G. Lucignani, G. Paganelli, E. Bombardieri, The use of standardized uptake values forassessing FDG uptake with PET in oncology: a clinical perspective, Nucl. Med. Commun. 25 (2004) 651–656.
[2] M. Okada, N. Sato, K. Ishii, K. Matsumura, M. Hosono, T. Murakami, FDG PET/CT versus CT, MR imaging and ^{67}Ga scintigraphy in the posttherapy evaluation of malignant lymphoma, RadioGraphics 30 (2010) 939–957.
[3] Martin S. Judenhofer, Simon R. Cherry, Applications for preclinical PET/MRI, Semin. Nucl. Med. 43 (1) (2013) 19–29.
[4] A. Takeda, N. Yokosuka, T. Ohashi, E. Kunieda, H. Fujii, Y. Aoki, N. Sanuki, N. Koike, Y. Ozawa, The maximum standardized uptake value (SUV_{max}) on FDG-PET is a strong predictor of local recurrence for localized non-small-cell lung cancer after stereotactic body radiotherapy (SBRT), Radiother. Oncol. 101 (2011) 291–297.
[5] H. Zhang, K. Wroblewski, S. Liao, R. Kampalath, B. Penney, Y. Zhang, Y.. Pu, Prognostic value of metabolic tumor burden from ^{18}F-FDG PET in surgical patients with non-small-cell lung cancer, Acad. Radiol. 20 (2013) 32–40.
[6] D Zhu, T Ma, Z Niu, J Zheng, A Han, S Zhao, et al., Prognostic significance of metabolic parameters measured by $^{(18)}$F-fluorodeoxyglucose positron emission tomography/computed tomography in patients with small cell lung cancer, Lung Cancer 73 (2011) 332–337.
[7] M.K. Chung, H.S. Jeong, S.G. Park, J.Y. Jang, Y.I. Son, J.Y. Choi, et al., Metabolic tumor volume of [^{18}F]-fluorodeoxyglucose positron emission tomography/computed tomography predicts short-term outcome to radiotherapy with or without chemotherapy in pharyngeal cancer, Clin Cancer Res. 15 (2009) 5861–5868.
[8] F Sampedro, A. Domenech, S Escalera, Obtaining quantitative global tumoral state indicators based on whole-body PET/CT scans: a breast cancer case study, Nucl. Med. Commun. 35 (4) (2014) 362–371.
[9] R. Arsanjani, Y. Xu, S Hayes, M. Fish, M. Lemley, J Gerlach, S Dorbala, S Berman, G Germano, P. Slomka, Comparison of fully automated computer analysis and visual scoring for detection of coronary artery disease from myocardial perfusion SPECT in a large population, J. Nucl. Med. 54 (2) (2013) 221–228.
[10] M. Hoffer, CT Teaching ManualGeorg Thieme Verlag (2000) 12–13.
[11] F. Sampedro, S. Escalera, A. Domenech, I. Carrió, Automatic tumor volume segmentation in whole-body PET/CT scans: a supervised learning approach, J. Med. Imaging Health Inf. 5 (2015) 1–10.
[12] A. Takeda, N. Yokosuka, T. Ohashi, E. Kunieda, H. Fujii, Y. Aoki, N. Sanuki, N. Koike, Y. Ozawa, The maximum standardized uptake value (SUV_{max}) on FDG-PET is a strong predictor of local recurrence for localized non-small-cell lung cancer after stereotactic body radiotherapy (SBRT), Radiother. Oncol. 101 (2011) 291–297.
[13] Wolpert, D. The supervised learning no-free-lunch theorems, in: Proceedings of the Sixth Online World Conference on Soft Computing in Industrial Applications, 2001.
[14] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Ann. Stat. 28 (2) (2000) 337–407.

[15] B. Boser, I. Guyon, V. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 144–152, 1992.
[16] W Chao, Y Chen, S Lin, Y Shih, S Tsang, Automatic segmentation of magnetic resonance images using a decision tree with spatial information, Comput. Med. Imaging Graph. 33 (2) (2009) 111–121.
[17] A Hain, J. Mao, K.. Mohiuddin, Artificial neural networks: a tutorial, computer —special issue: neural computing: companion issue to Spring 1996, IEEE Comput. Sci. Eng. Arch. 29 (3) (1996) 31–44.
[18] Oriol Pujol Sergio Escalera, Petia Radeva, On the decoding process in ternary error-correcting output codes, Trans. Pattern Anal. Mach. Intell., 32, 2010120–134 (ISSN 0162-8828).
[19] Guan, H., Kubota, T., Huang, X. Sean, X., Turk, M. Automatic hot spot detection and segmentation in whole body FDG-PET images, in: Proceedings of the International Conference on Image Processing, October 8–11, 2006, pp 85–88.
[20] TD Cannon, Y Chung, G He, D Sun, A Jacobson, TG van Erp, et al., Progressive reduction in cortical thickness as psychosis develops: a multisite longitudinal neuroimaging study of youth at elevated clinical risk, Biol. Psychiatry (Jun) (2014), http://dx.doi.org/10.1016/j.biopsych.2014.05.023.
[21] R.D. Rudyanto, G. Bastarrika, G. de Biurrun, J. Agorreta, L.M. Montuenga, C. Ortiz-de-Solorzano, A. Muñoz-Barrutia, Individual nodule tracking in micro-CT images of a longitudinal lung cancer mouse model, Med. Image Anal. 17 (8) (2013) 1095–1105. http://dx.doi.org/10.1016/j.media.2013.07.002.
[22] SM LandauM.A. MintunA.D. Joshi, R.A. Koeppe, R.C. Petersen, P.S. Aisen, M.W. Weiner, W.J. Jagust, Alzheimer's disease neuroimaging initiative. Amyloid deposition, hypometabolism, and longitudinal cognitive decline, Ann. Neurol. 72 (4) (2012) 578–586. http://dx.doi.org/10.1002/ana.23650.
[23] Syamsiah B.T. Mashohor Afsaneh Jalalian, Hajjah Rozi Mahmud, M. Iqbal, B. Saripan, Abdul Rahman B. Ramli, Babak Karasfi, Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review, Clin. Imaging 37 (3) (2013) 420–426. http://dx.doi.org/10.1016/j.clinimag.2012.09.024 (ISSN 0899-7071).
[24] J. Ramírez, J.M. Górriz, D. Salas-Gonzalez, A. Romero, M. López, I. Álvarez, M. Gómez-Río, Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features, Inf. Sci. 237 (July) (2013) 59–72. http://dx.doi.org/ 10.1016/j.ins.2009.05.012 (ISSN 0020-0255).

Frederic Sampedro, obtained the Computer Science B.Sc., with honors in 2010 from the University of Barcelona (UB). In 2012, he obtained a M.Sc., degree in Electrical Engineering and a M.Sc., degree in Biomedical Engineering from the University of Barcelona and the Polytechnic University of Catalonia, respectively. He is currently working in his Ph.D. thesis about automatic tumor quantification in PET/CT imaging.

Sergio Escalera, obtained the Ph.D. degree on Multi-class visual categorization systems at Computer Vision Center (Universitat Autonoma de Barcelona) obtaining the 2008 best Thesis award on Computer Science. He leads the Human Pose Recovery and Behavior Analysis Group. He is a lecturer of the Department of Applied Mathematics and Analysis, Universitat de Barcelona. He is Editor-in-Chief of American Journal of Intelligent Systems. He is advisor of ChaLearn Challenges in Machine Learning. He is member of the AERFAI Spanish Association on Pattern Recognition and ACIA Catalan Association of Artificial Intelligence.

Anna Domenech, is a nuclear medicine physician at the Hospital de Sant Pau (Barcelona, Spain). She obtained the B.Sc., and Ph.D. degrees in Medicine from the Universitat Autonoma de Barcelona and the University of Barcelona in 2003 and 2011, respectively. Her research interests include the clinical role of sentinel lymph node examination and tumor quantification strategies in PET/CT imaging.

Ignasi Carrio, MD Ph.D. FEBNM FESC FRCP is a full professor of nuclear medicine at the Universitat Autonoma de Barcelona and the director of the Nuclear Medicine department at the Hospital de Sant Pau (Barcelona, Spain). He was the former president (2004-2006) of the European Association of Nuclear Medicine and is the Editor-in-Chief of the European Journal of Nuclear Medicine and Molecular Imaging.

Original article

Nuclear Medicine Communications

Deriving global quantitative tumor response parameters from ^{18}F-FDG PET-CT scans in patients with non-Hodgkin's lymphoma

Frederic Sampedro[a], Anna Domenech[c], Sergio Escalera[b,d] and Ignasi Carrió[c]

Objectives **The aim of the study was to address the need for quantifying the global cancer time evolution magnitude from a pair of time-consecutive positron emission tomography-computed tomography (PET-CT) scans. In particular, we focus on the computation of indicators using image-processing techniques that seek to model non-Hodgkin's lymphoma (NHL) progression or response severity.**

Materials and methods **A total of 89 pairs of time-consecutive PET-CT scans from NHL patients were stored in a nuclear medicine station for subsequent analysis. These were classified by a consensus of nuclear medicine physicians into progressions, partial responses, mixed responses, complete responses, and relapses. The cases of each group were ordered by magnitude following visual analysis. Thereafter, a set of quantitative indicators designed to model the cancer evolution magnitude within each group were computed using semiautomatic and automatic image-processing techniques. Performance evaluation of the proposed indicators was measured by a correlation analysis with the expert-based visual analysis.**

Results **The set of proposed indicators achieved Pearson's correlation results in each group with respect to the expert-based visual analysis: 80.2% in progressions, 77.1% in partial response, 68.3% in mixed response, 88.5% in complete response, and 100% in relapse. In the progression and mixed response groups, the proposed indicators outperformed the common indicators used in clinical practice [changes in metabolic tumor volume, mean, maximum, peak standardized uptake value (SUV_{mean}, SUV_{max}, SUV_{peak}), and total lesion glycolysis] by more than 40%.**

Conclusion **Computing global indicators of NHL response using PET-CT imaging techniques offers a strong correlation with the associated expert-based visual analysis, motivating the future incorporation of such quantitative and highly observer-independent indicators in oncological decision making or treatment response evaluation scenarios.** ***Nucl Med Commun*** **36:328–333**

Nuclear Medicine Communications 2015, **36**:328–333

Keywords: image analysis, non-Hodgkin's lymphoma, PET-computed tomography, tumor response

[a]Department of Radiology, Faculty of Medicine, [b]Computer Vision Center, Building O, Autonomous University of Barcelona, [c]Department of Nuclear Medicine, Hospital de Sant Pau and [d]Department of Applied Mathematics, Faculty of Mathematics, University of Barcelona, Barcelona, Spain

Correspondence to Frederic Sampedro, MSc, Faculty of Medicine, Autonomous University of Barcelona, 08193 Barcelona, Spain
Tel: +34 699 805 231; e-mail: fredsampedro@gmail.com

Received 5 September 2014 **Revised** 19 October 2014
Accepted 26 November 2014

Introduction

Fluorine-18 fluorodeoxyglucose (^{18}F-FDG) positron emission tomography-computed tomography (PET-CT) has become a standard imaging method for the time monitoring of treatment response in a variety of tumors [1–3]. From a pair of time-consecutive whole-body PET-CT scans nuclear medicine physicians assess a patient's cancer progression or response condition following a trained visual and semiquantitative analysis of both images. Thereafter, generally, a categorical and qualitative diagnosis is provided, such as 'good response', 'slight progression', or 'strong relapse'. Although this type of information is generally enough in the clinical routine, it lacks observer independence and does not provide a continuous response scale to accurately compare between cases.

In this work we address the need and computation of observer-independent global quantitative tumor response indicators from a pair of time-consecutive PET-CT scans. This complementary information to the physician's visual analysis would prove especially useful in comprehensive oncological treatment response evaluation and comparison scenarios, as well as in the context of studying possible cancer evolution differences related to particular clinical profiles.

This issue has been partially addressed in the literature in the form of relating time changes in local tumor metabolic activity or volume with surgical outcome parameters [4–8]. Although this methodology is well suited to recognize the value of quantifying PET-CT images, it does not provide a sound framework for designing and evaluating the proposed global response indicators due to several reasons.

First, changes in cancer spread are not taken into consideration, which, as derived from Sampedro *et al.* [9] and described later in this paper, play a key role in measuring the cancer progression or response magnitude. Second, it

DOI: 10.1097/MNM.0000000000000256

is important to note, in the general case, the lack of a well-defined gold standard indicator to compare the proposed global response indicators. In particular, non-Hodgkin's lymphoma (NHL) response or progression magnitude is not well described by any clinical continuous parameter. Even if the international prognostic index is considered the current prognostication system for NHL, prognostic heterogeneity is suggested to exist among the patients within the same international prognostic index risk group [9–11]. In such a scenario, the performance of the proposed indicators can be addressed by using the information resulting from an expert-based ordering by magnitude of the cases where the indicator performance is to be measured, as shown in [3].

Thus, in this work, we start from an ordered set of NHL response/progression cases based on its magnitude (derived by the visual analysis of a consensus of experts focusing on time changes in tumor volume, aggressiveness, and spread). Then, from its associated pair of PET-CT scans, we propose and compute a set of global response/progression indicators by quantifying time changes in the segmented metabolic tumor volumes (also provided by nuclear medicine physicians). Indicator performance is addressed by a correlation analysis with the initial expert-based ordering. Aiming to maximize observer independence in the indicator computations, the possibility of using completely automatic PET tumor volume segmentation techniques is also addressed.

Materials and methods

A set of 178 whole-body FDG-PET/CT scans corresponding to NHL lymphoma patients were acquired from the Phillips Nuclear Medicine workstation at Hospital de Sant Pau (Barcelona, Spain) following all international PET/CT imaging acquisition protocols [12]. From its digital imaging and communications in medicine (DICOM) files, two coregistered three-dimensional volumes were obtained for each scan: a PET volume, in standardized uptake value (SUV) [13] units, and a CT volume (in Hounsfield units) [14]. They corresponded to 89 pairs of time-consecutive scans of the patients. The time elapsed between scans varied depending on the clinical management of each patient, with a median of 3.2 months and an interquartile range of 2 months.

Classification of each cancer evolution condition was carried out by a consensus of three independent nuclear medicine physicians into progression (31), partial response (28), mixed response (nine), complete response (13), or relapse (eight). The classification criteria were based on changes in tumor volume, aggressiveness (represented by its metabolic activity through its SUV), and spread, as illustrated in Fig. 1. Figures 2 and 3 show examples of real cases of each cancer evolution condition.

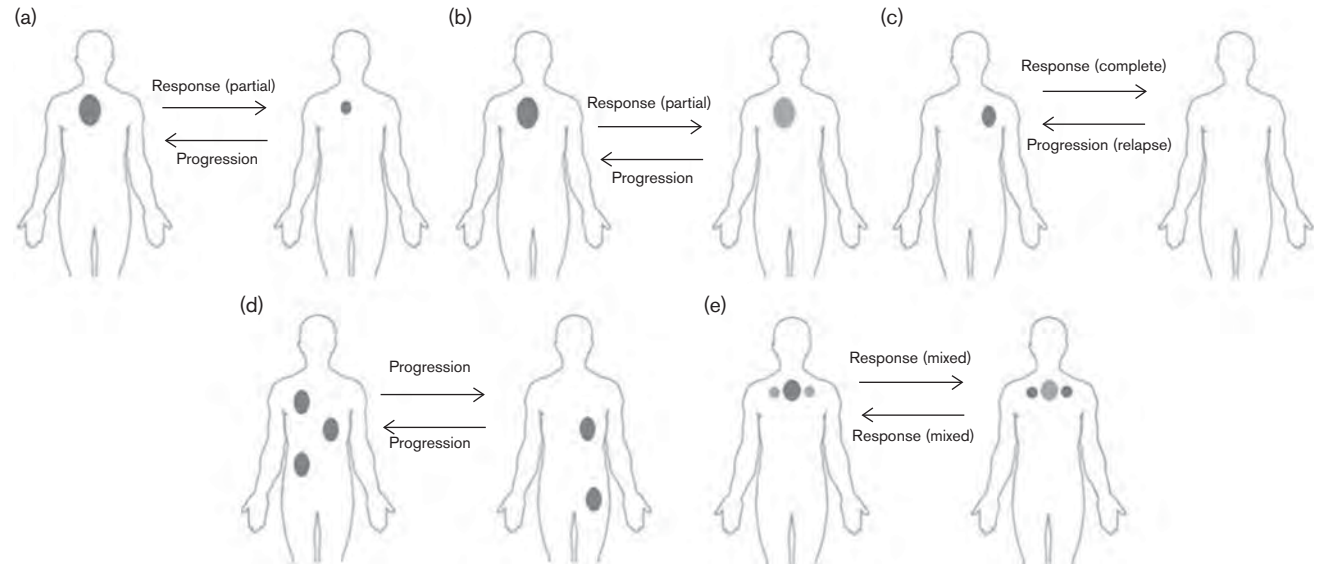
Note that the cases illustrated in Fig. 1 represent the canonical cancer evolution conditions; that is, in practice, real cases may be combinations of those cases. For instance, a progression case can be presented both with an increase in tumor volume (or uptake) and with the appearance of new lesions, a response case with both a decrease in volume and uptake, or a mixed response case with both increases and decreases in volume and uptake of the persisting tumor lesions.

Then, the cases of each group are ordered by its magnitude according to the following visual criteria. For progression cases, relative increases in volume or aggressiveness in the existing tumor lesions are considered less severe than the appearance of new tumor lesions in adjacent or distant anatomical locations, respectively. However, all these variables interact, in the sense that strong volume increases of existing lesions may be considered more severe than the sole appearance of small adjacent new lesions. The ordering of partial responses is analogous, but considering the relative volume, aggressiveness, and spread it decreases (emphasizing the global tumor size reduction). From an imaging point of view, the ordering of relapses and complete responses is analogous to that of progressions and partial responses without any tumor presence in one of the scans. Mixed responses are ordered considering the overall balance of tumor volume and uptake increases and decreases of the existing tumor lesions. Figure 3 shows an ordering example of a subset of the progressions and partial responses considered in this study.

The main goal of this work was to analyze the best global quantitative indicators that model each of the cancer evolution groups so as to obtain a continuous analog of the visual qualitative assessment. The performance of each proposed indicator will be addressed by comparing (using the Pearson correlation coefficient) the ordering provided by the medical experts with the order obtained by the indicator of the same cases.

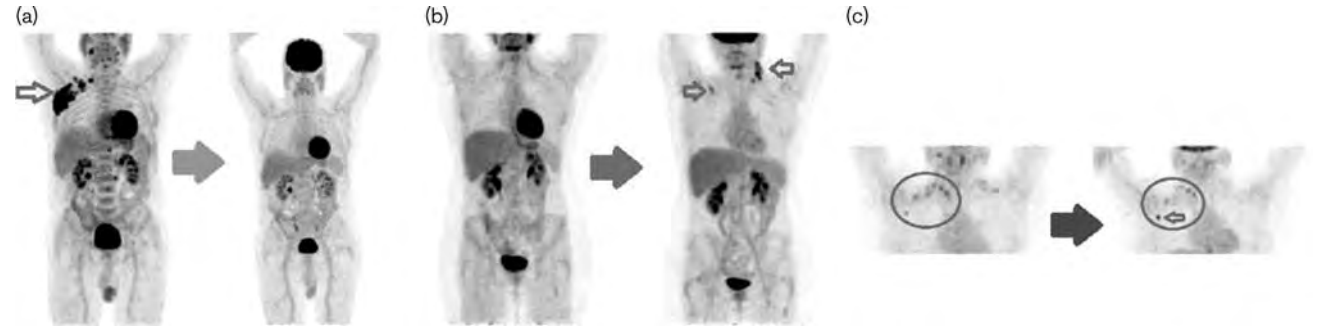
On the design of such global indicators, the ones more commonly used in clinical practice are first considered. Conceptually, in the presence of more than a single tumor lesion or highly heterogenous tumor tissue (e.g. the presence of necrotic tissue in any of the scans), global changes in SUV_{mean}, SUV_{max}, or SUV_{peak} [4,13,15,16] will not appropriately model the strength of the progression or response condition, as they are unable to model volume increases or the appearance of new tumor lesions. In contrast, global changes in whole-body metabolic tumor volume (WBMTV) or total lesion glycolysis (TLG) [15] offer a better overall description of the magnitude of the cancer evolution. However, they still suffer from conceptual limitations: consider the cases modeled in Fig. 1d and in particular the top right progression case in Fig. 3. In such a case, these indicators may not even be valuable, as the global WBMTV has in fact decreased in time while the case is considered a strong cancer progression.

Fig. 1



Illustrations of the typical cancer evolution scenarios in nuclear medicine. When a single tumor lesion is present, its change in volume or intensity in time defines its progression or response condition (a–c). In the multilesion case, the spread of the cancer into new anatomical locations (regardless of the volume change) is associated with a progression scenario (d), whereas the intensity increase in any of the lesions is clinically associated with a mixed response scenario (e). Intensity of the tumor lesion is represented by its grayscale intensity.

Fig. 2

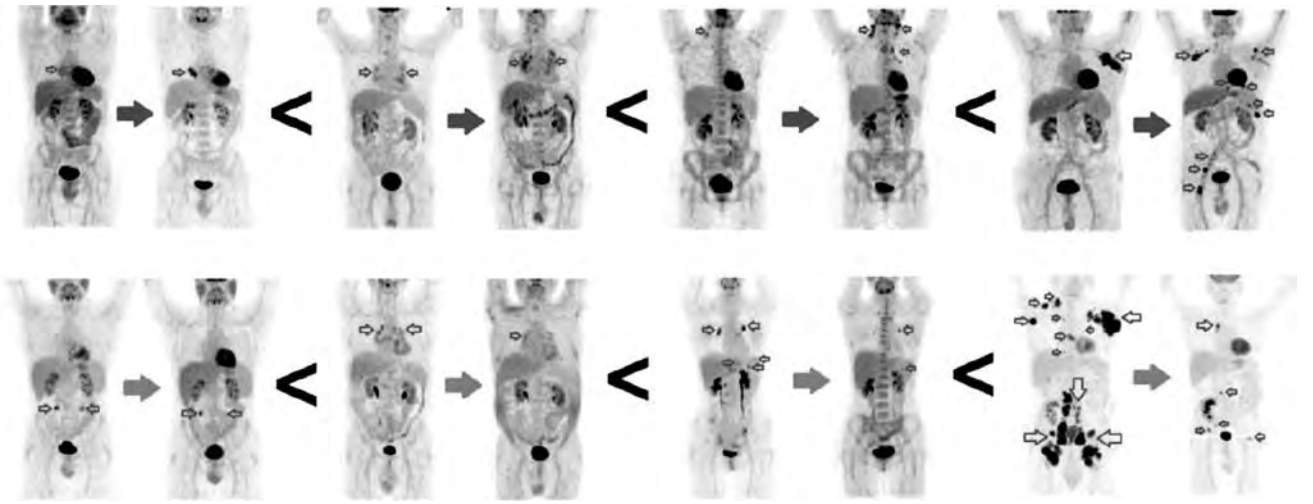


Clinical examples of complete response (a), relapse (b), and mixed response (c). The thick arrows represent the direction of time. Each PET scan is visualized using its maximum intensity projection.

Therefore, we propose a new set of indicators that seek to model more accurately the global progression or response magnitude. With an eye to future technological advances, we focus only on quantitative indicators that can be computed from the PET three-dimensional tumor segmentation masks of both time-consecutive scans. These, in the future, may be obtained accurately in an automatic manner using recent advances in machine learning-based segmentation techniques [17], thus obtaining full observer independence in the whole process. Nevertheless, as current automatic segmentation methods do not achieve the required accuracy to compute reliable indicators in this scenario [17], we focus on the use of expert-guided semiautomatic tumor segmentation masks that, although introducing a slight observer dependence and a highly time-consuming step, provide accurate and reliable estimators of the underlying phenomena. Figure 4 illustrates this reasoning.

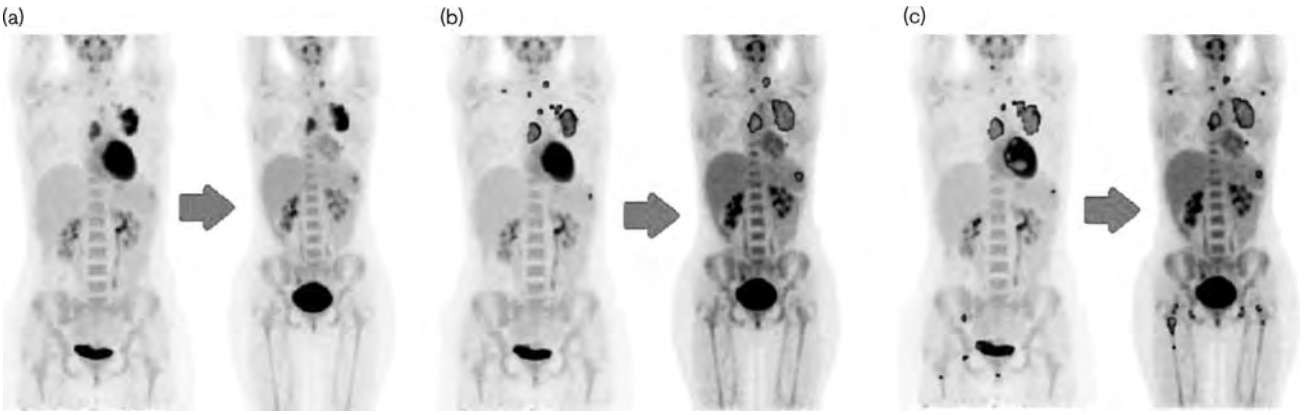
The key clinical variable that the new set of indicators need to model are changes in cancer spread, which are not modeled by the common indicators described before. A first piece of information in this respect is the change in the number of tumor-related lesions. These can be modeled computationally by the change in the number of connected components (ΔNCC) between the pair of PET tumor segmentation masks, as from an image-processing point of view a connected component [18]

Fig. 3



Ordering of a subset of progression (top) and partial response (bottom) cases of this study.

Fig. 4



Cancer progression example (a) and its associated expert-guided (b) and automatic (c) three-dimensional tumor segmentation masks. Note how the presence of errors in automatic segmentation masks may lead to conclude about a false global decrease in WBTMV (c) in an actual volume increase scenario (b). WBTMV, whole-body metabolic tumor volume.

in the tumor segmentation mask can be associated with a single tumor-related lesion [19]. Although clearly a high ΔNCC in magnitude will be likely associated with the strength of the cancer evolution, this parameter suffers from noisy behavior due to possible segmentation inaccuracies [19] and does not quantify the actual volume of the new tumor lesions. Furthermore, it will not recognize the cancer progression scenario illustrated in Fig. 1d, in which, even though a decrease in NCC (i.e. number of tumor lesions) is observed, an underlying progression condition could be present.

To overcome those limitations, another clinically relevant parameter is computed, denoted as V_N. V_N is designed to quantify the amount of new tumor volume that appeared in the second scan with respect to the first. V_N does not include volume increases of the existing lesions; that is, it only adds up the volume of tumor lesions that appeared in new anatomical locations, thus quantifying the spread strength. Note that this indicator will effectively recognize and quantify the progression cases illustrated in Fig. 1d. The computation of V_N from the pair of time-consecutive PET tumor segmentation masks is nontrivial and described in [20]. In short, both PET scans are realigned and new tumor lesions are detected and quantified from the subtraction of the realigned segmentation masks.

Also, as has been mentioned, the appearance of tumor lesions in new organs or distant anatomical locations is considered to worsen the cancer progression condition.

To model this effect, we introduce the number of significantly new tumor lesions (nSNTL) parameter and approximate it computationally in the following manner. As the set of new tumor lesions of the tumor segmentation mask from the second scan is obtained during the computation of V_N, the remaining task is to identify and count which of those lesions (i.e. connected components in the mask) can be classified as belonging to a new organ or being sufficiently distant from the lesions of the first scan. For that, one of these two conditions must hold: either the mean Hounsfield unit value of a given lesion is significantly different ($P < 0.05$) from that of all the lesions in the first scan, or it is significantly distant (>1% of the patient's body surface area [21]) from them.

Finally, an indicator that aids in quantifying the magnitude of mixed responses is presented, denoted as A_N. A_N is designed to quantify the amount of tumor volume that increased its activity by more than 20% in the second scan relative to the amount of tumor volume in the first scan. Again, the computation of A_N from the pair of tumor segmentations masks is nondirect and described [20].

Results

Table 1 shows the performance results, in this context, of the common indicators used in clinical practice. Strong correlations are only observed in partial response, relapse, and complete response cases.

Table 2 shows the performance results obtained by combining them with the proposed alternative indicators described in the previous section. Substantial performance increases are shown in progression and mixed response cases. No strong correlation was observed in any scenario if using automatic segmentation procedures.

Finally, although the indicators presented in Table 2 are the ones that obtained the best performance results on our particular data set, very similar results were obtained when using ΔTLG instead of ΔWBMTV (79.6% correlation in progression cases, 76.7% in partial responses, and 90.5% in relapses). Also, the complete response cases were also modeled accurately by ΔWBMTV and ΔTLG (both showing 83.5% correlation).

Table 1 Pearson's correlation results of the common indicators used in clinical practice with respect to the expert-based visual ordering

Correlation (%)	ΔWBMTV	ΔSUV_{mean}	ΔSUV_{max}	ΔSUV_{peak}	ΔTLG
Progression	32.3	5.7	4.2	13.8	30.7
Partial response	76.9	48.1	59.6	43.7	73.8
Mixed response	8.3	20.0	13.3	26.7	25.0
Relapse	100	54.8	90.5	78.6	90.5
Complete response	88.5	44.0	64.8	80.8	83.5

max, maximum; SUV, standardized uptake value; TLG, total lesion glycolysis; WBMTV, whole-body metabolic tumor volume.
Δ Relative change.

Discussion

Several noteworthy conclusions can be drawn from our results. First, note that the conceptual limitations of this set of indicators described in the previous section are empirically observed in our data set. Second, the formulas of the indicators that best model the cases in the data set are a highly consistent mathematical representation of the physician global visual analysis criteria. Third, a substantial performance increase is shown in the progression and mixed response scenarios with respect to the indicators in Table 1, which demonstrates the relevance of the new set of proposed indicators in modeling real NHL evolution cases. Also, the stable results observed in the rest of the scenarios are also coherent, as the change in the overall tumor size and extension (modeled by ΔWBMTV or including the tumor activity information using TLG) is clearly the most important visual criterion in those cases. Fourth, the most difficult (e.g. the one with more discrepancies in the consensus of physicians performing the visual analysis) NHL evolution scenario to order by magnitude was the mixed response, which also showed the worse indicator correlation results. Finally, fifth, as mentioned in the Materials and methods section, current completely automatic tumor segmentation techniques are not capable of offering reliable parameters in this clinical context.

We also considered the possibility of including the time elapsed between scans as another factor in the indicator formulas, as clearly the same cancer progression or response could be considered 'stronger' if it was produced in a shorter period of time. However, we consider that the evaluation of this parameter, in conjunction with other clinical variables such as the specific treatment design of each patient, should be carried out at the oncological management level and not included in the nuclear medicine PET/CT diagnostic quantification framework. Similarly, we only considered mathematical combinations of the proposed indicators that had a sensible clinical basis, and left as future work a possible in-depth analysis on fitting parametrical statistical models to the proposed combined indicator formulas to study the possible asymmetric weight distribution of each indicator.

Finally, we consider that the incorporation of this type of quantitative parameters in nuclear medicine diagnostic frameworks could increase its overall potential. However, a large amount of future work remains. On one hand, expert-guided semiautomatic segmentation of whole-body PET scans is a highly time-consuming task and therefore is typically unfeasible in the clinical routine. In this work we showed that current completely automatic segmentation techniques are unable to provide reliable indicators in this diagnostic context, motivating the initiation of further research in this area. In contrast, the incorporation of this type of indicators at the oncological management level would require a previous in-depth

Table 2 Pearson's correlation results for the indicators that obtained the best performance results on the data set

	Indicator	Correlation (%)	Correlation (%): automatic segmentation
Progression	$\Delta WBMTV \times V_N \times nSNTL$	80.2	18.1
Partial response	$\Delta WBMTV \times (1 + \lvert\Delta NCC\rvert)$	77.1	32.2
Mixed response	$A_N/\Delta WBMTV$	68.3	28.3
Relapse	$\Delta WBMTV$	100	45.2
Complete response	ΔNCC	88.5	41.2

Note that V_N and nSNTL are only included in the product if they have a value greater than zero. Results are also reported using completely automatic segmentation techniques.
NCC, number of connected components; nSNTL, number of significantly new tumor lesions; WBMTV, whole-body metabolic tumor volume.

analysis of its exact role as well as its possible limitations in the clinical context, including its performance evaluation within alternative gold standard frameworks.

Conclusion

Addressing the need for obtaining a global continuous and observer-independent representation of the cancer evolution magnitude from a pair of whole-body PET-CT scans, in this work we proposed a set of global indicators of NHL response computed through imaging techniques that offered strong correlation results with the associated expert-based visual analysis.

Acknowledgements

The work of Frederic Sampedro is supported by the Spanish government FPU (Formación del Profesorado Universitario) doctoral grant and Spanish Project TIN2013-43478-P.

Conflicts of interest

There are no conflicts of interest.

References

1 Ciernik IF, Dizendorf E, Baumert BG, Reiner B, Burger C, Davis JB, *et al.* Radiation treatment planning with an integrated positron emission and computer tomography (PET/CT): a feasibility study. *Int J Radiat Oncol Biol Phys* 2003; **57**:853–863.
2 Rosenman J. Incorporating functional imaging information into radiation treatment. *Semin Radiat Oncol* 2001; **11**:83–92.
3 Czernin J, Phelps ME. Positron emission tomography scanning: current and future applications. *Annu Rev Med* 2002; **53**:89–112.
4 Takeda A, Yokosuka N, Ohashi T, Kunieda E, Fujii H, Aoki Y, *et al.* The maximum standardized uptake value (SUV_{max}) on FDG-PET is a strong predictor of local recurrence for localized non-small-cell lung cancer after stereotactic body radiotherapy (SBRT). *Radiother Oncol* 2011; **101**:291–297.
5 Zhang H, Wroblewski K, Liao S, Kampalath R, Penney BC, Zhang Y, Pu Y. Prognostic value of metabolic tumor burden from (18)F-FDG PET in surgical patients with non-small-cell lung cancer. *Acad Radiol*, **20**:32–402013.
6 Kim MH, Lee JS, Mok JH, Lee K, Kim KU, Park HK, *et al.* Metabolic burden measured by (18)F-fluorodeoxyglucose positron emission tomography/computed tomography is a prognostic factor in patients with small cell lung cancer. *Cancer Res Treat* 2014; **46**:165–171.
7 Sridhar P, Mercier G, Tan J, Truong MT, Daly B, Subramaniam RM. FDG PET metabolic tumor volume segmentation and pathologic volume of primary human solid tumors. *Am J Roentgenol* 2014; **202**:1114–1119.
8 Obara P, Pu Y. Prognostic value of metabolic tumor burden in lung cancer. *Chin J Cancer Res* 2013; **25**:615–622.
9 Sampedro F, Domenech A, Escalera S. Obtaining quantitative global tumoral state indicators based on whole-body PET/CT scans: a breast cancer case study. *Nucl Med Commun* 2014; **35**:362–371.
10 Jung SH, Yang DH, Ahn JS, Kim YK, Kim HJ, Lee JJ. Serum lactate dehydrogenase with a systemic inflammation score is useful for predicting response and survival in patients with newly diagnosed diffuse large b-cell lymphoma. *Acta Haematol* 2014; **133**:10–17.
11 Hermans J, Krol AD, van Groningen K, Kluin PM, Kluin-Nelemans JC, Kramer MH, *et al.* International prognostic index for aggressive non-Hodgkin's lymphoma is valid for all malignancy grades. *Blood* 1995; **86**:1460–1463.
12 Park JH, Yoon DH, Kim DY, Kim S, Seo S, Jeong Y, *et al.* The highest prognostic impact of LDH among international prognostic indices (IPIs): an explorative study of five IPI factors among patients with DLBCL in the era of rituximab. *Ann Hematol* 2014; **93**:1755–1764.
13 Sampedro F, Escalera S, Domenech A, Carrió I. A computational framework for cancer response assessment based on oncological PET-CT scans. *Comput Biol Med* 2014; **55**:92–99.
14 Hoffer M. *CT teaching manual*. Leipzig, Germany: Georg Thieme Verlag; 2000. pp. 12–13.
15 Vanderhoek M, Perlman SB, Jeraj R. Impact of the definition of peak standardized uptake value on quantification of treatment response. *J Nucl Med* 2012; **53**:4–11.
16 Chan WK, Mak HK, Huang B, Yeung DW, Kwong DL, Khong PL, *et al.* Nasopharyngeal carcinoma: relationship between 18F-FDG PET-CT maximum standardized uptake value, metabolic tumour volume and total lesion glycolysis and TNM classification. *Nucl Med Commun* 2010; **31**:206–210.
17 Achury C, Estorch M, Domènech A, Camacho V, Flotats A, Jaller R, *et al.* Interpretation of thyroid incidentalomas in (18)F-FDG PET/CT studies. *Rev Esp Med Nucl Imagen Mol* 2014; **33**:205–209.
18 Biancardi A, Segovia M. Adaptive segmentation of MR axial brain images using connected components. *Lect Notes Comput Sci* 2001; **2059**:295–302.
19 Byun BH, Kong CB, Park J, Seo Y, Lim I, Choi CW, *et al.* Initial metabolic tumor volume measured by ^{18}F-FDG PET/CT can predict the outcome of osteosarcoma of the extremities. *J Nucl Med* 2013; **54**:1725–1732.
20 Sampedro F, Escalera S, Domenech A, Carrió I. Automatic tumor volume segmentation in whole-body PET/CT scans: a supervised learning approach. *J Med Imaging Health Infor* 2015; **5**:1–10.
21 Mosteller RD. Simplified calculation of body-surface area. *N Engl J Med* 1987; **317**:1098.

Rev Esp Med Nucl Imagen Mol. 2016;xxx(xx):xxx–xxx

Original article

Computing quantitative indicators of structural renal damage in pediatric DMSA scans

F. Sampedro[a,*], A. Domenech[b], S. Escalera[c,d], I. Carrio[b]

[a] Autonomous University of Barcelona, Faculty of Medicine, 08193 Barcelona, Spain
[b] Hospital de Sant Pau, Nuclear Medicine Department, Carrer Sant Quintí 89, 08026 Barcelona, Spain
[c] Computer Vision Center, Campus UAB, Edifici O, 08193 Bellaterra, Barcelona, Spain
[d] Department of Applied Mathematics and Analysis, Universitat de Barcelona, Gran Via de les Corts 585, 08007 Barcelona, Spain

ARTICLE INFO

Article history:
Received 21 April 2016
Accepted 17 June 2016
Available online xxx

Keywords:
DMSA scan
Renal damage
Image processing
Quantitative analysis

ABSTRACT

Objectives: The proposal and implementation of a computational framework for the quantification of structural renal damage from ^{99m}Tc-dimercaptosuccinic acid (DMSA) scans.
The aim of this work is to propose, implement, and validate a computational framework for the quantification of structural renal damage from DMSA scans and in an observer-independent manner.
Materials and methods: From a set of 16 pediatric DMSA-positive scans and 16 matched controls and using both expert-guided and automatic approaches, a set of image-derived quantitative indicators was computed based on the relative size, intensity and histogram distribution of the lesion. A correlation analysis was conducted in order to investigate the association of these indicators with other clinical data of interest in this scenario, including C-reactive protein (CRP), white cell count, vesicoureteral reflux, fever, relative perfusion, and the presence of renal sequelae in a 6-month follow-up DMSA scan.
Results: A fully automatic lesion detection and segmentation system was able to successfully classify DMSA-positive from negative scans (AUC = 0.92, sensitivity = 81% and specificity = 94%). The image-computed relative size of the lesion correlated with the presence of fever and CRP levels (p <0.05), and a measurement derived from the distribution histogram of the lesion obtained significant performance results in the detection of permanent renal damage (AUC = 0.86, sensitivity = 100% and specificity = 75%).
Conclusions: The proposal and implementation of a computational framework for the quantification of structural renal damage from DMSA scans showed a promising potential to complement visual diagnosis and non-imaging indicators.

Cómputo de indicadores cuantitativos de daño renal estructural en imágenes DMSA pediátricas

Palabras clave:
DMSA
Daño renal
Análisis de imagen
Análisis cuantitativo

RESUMEN

Objetivos: En el presente trabajo se propone, implementa y valida un entorno computacional de cuantificación de imágenes con ^{99m}Tc-ácido dimercaptosuccínico (DMSA) con el objetivo de obtener indicadores cuantitativos del daño renal subyacente. Estos indicadores se validan en un contexto de imágenes DMSA pediátricas, dada su relevancia en el diagnóstico de pielonefritis aguda y cicatrices renales.
Materiales y métodos: Partiendo de un conjunto de 16 imágenes DMSA positivas para daño renal y 16 controles apareados por edad y sexo, se proponen y calculan una serie de indicadores cuantitativos basados en el área relativa lesionada y la distribución de su histograma. Se implementan aproximaciones manuales y automáticas para dicho cómputo. Los indicadores obtenidos se correlacionan con otras variables clínicas de interés en este contexto, como la proteína C reactiva, la cuenta leucocitaria, el reflujo vesicouretral, la fiebre, la perfusión relativa, y la presencia de secuelas renales en la imagen DMSA a los 6 meses de seguimiento.
Resultados: El sistema implementado de detección y cuantificación de lesiones renales obtuvo un rendimiento significativo discriminando las imágenes DMSA positivas de las negativas (AUC = 0,92, sensibilidad = 81% y especificidad = 94%). El indicador de área relativa de la lesión correlacionó con los niveles de proteína C reactiva y la presencia de fiebre (p<0,05). Finalmente, un indicador derivado de las propiedades del histograma de la lesión obtuvo un rendimiento significativo en la detección de la presencia de secuelas renales (AUC = 0,86, sensibilidad = 100% y especificidad = 75%).

* Corresponding author.
E-mail address: fredsampedro@gmail.com (F. Sampedro).

http://dx.doi.org/10.1016/j.remn.2016.06.010

Conclusiones: La propuesta e implementación de un entorno computacional para la obtención de indicadores cuantitativos a partir de imágenes DMSA muestra un potencial prometedor para complementar el diagnóstico visual.

Introduction

^{99m}Tc-dimercaptosuccinic acid (DMSA) scans are a valuable nuclear medicine test in assessing renal morphology and structural damage. At the time of this writing, planar-image DMSA is the gold standard for the diagnosis of acute pyelonephritis and renal scars.[1,2] This particular clinical context is especially relevant within the pediatric population.[3–5]

An important limitation of this technique is that it does not distinguish accurately lesions that will spontaneously resolve from those which will cause permanent renal damage.[6] For that, a 6-month follow-up DMSA scan is needed in order to confirm a renal scar diagnosis, representing a major limitation both in clinical and economic terms.

DMSA scan evaluation by the trained physician remains purely visual. As in most medical imaging scenarios, while this approach is accurate enough in many cases, it suffers from inter- and intra-observer variabilities.[7,8] Additionally, its diagnostic product is descriptive and categorical, lacking a continuous modeling of the underlying renal damage.

With computational advances, the trend to try to complement the visual diagnostic products with image-derived quantitative and observer-independent parameters is spreading in the field. Although quantitative DMSA image analysis has been used in the field,[9,25,26] to the best of our knowledge, the computation of image-derived quantitative indicators of structural renal damage in DMSA scans has not been addressed by means of a comprehensive computational framework.

In this work, we propose for the first time a DMSA segmentation and quantification framework that seeks to provide clinically valuable indicators for the assessment of structural renal damage. In particular, we aim to compute image-derived quantitative and subject-independent parameters designed to model accurately the underlying renal pathophysiology observed in DMSA scans. The performance such indicators will be evaluated within three different contexts: automatic renal damage detection, indicators' correlation with non-imaging clinical data and early permanent renal lesion detection.

Materials and methods

Demographics and DMSA acquisition

A total of 16 pediatric DMSA scans with visually-diagnosed structural kidney damage, its 6-month follow-up DMSA scans, and 16 age- and sex-matched controls were obtained from the Philips Precedence workstation at the Nuclear Medicine department of Hospital de Sant Pau, Barcelona, Spain.

Table 1
Demographic and clinical data for the DMSA-positive group. Detailed information about the clinical relevance of these variables can be found in Ref. 10.

	Proportion of subjects with the parameter available	Mean ± standard deviation or proportion
Age (months)	16/16	27 ± 32
Sex (male/female)	16/16	12/16 females
Weight (kg)	16/16	10.8 ± 6.3
Fever (positive > 38 °C)	15/16	13/15 positive
C reactive protein (CRP) (mg/L)	15/16	112.5 ± 84.1
Leukocyte count (mil/mm^3)	14/16	16.1 ± 7.6
Vesicoureteral reflux (positive/negative)	14/16	1/14 positive
Relative perfusion of the affected kidney (%)	16/16	47.6 ± 5.5
Chronic lesion, based on 6-month DMSA follow-up (positive/negative)	16/16	4/16 chronic lesion

The 16 pathological DMSA scans showed clearly identifiable upper-pole single-kidney lesions, which are the most prevalent in our center. The image type of choice to be analyzed in this work is the white-background posterior projection of the DMSA acquisition.

Several demographic and clinical variables of interest in this context were also obtained for the 16 pathological cases (Table 1).

The 16 DMSA positive scans are intended to represent a wide spectrum of renal damage scenarios, making them a valuable set to test the efficiency of the proposed quantitative indicators at modeling the underlying renal pathology (Fig. 1).

Computation of DMSA-derived quantitative indicators

In order to quantify the structural kidney damage (SKD) within the DMSA scans, two approaches (manual and automatic) were conducted regarding the image segmentation of the pathological areas. Then, from the obtained lesion's segmentation, a set of image-derived quantitative indicators is computed.

Manual segmentation and quantification methodology

An expert-guided manual segmentation framework was custom-build using Matlab®.[11] Given the low resolution of the

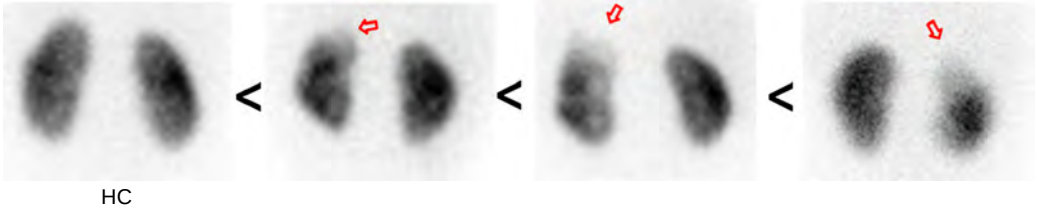


Fig. 1. Four examples of DMSA scans illustrating the pathological spectrum to be modeled by the proposed quantitative indicators. HC: healthy control. The patient of the third scan suffered renal sequelae as opposed to the patients of the other two pathological scans. The illustrative conceptual ordering of the positive scans was based on visual inspection of the damaged area.

Please cite this article in press as: Sampedro F, et al. Computing quantitative indicators of structural renal damage in pediatric DMSA scans. Rev Esp Med Nucl Imagen Mol. 2016. http://dx.doi.org/10.1016/j.remn.2016.06.010

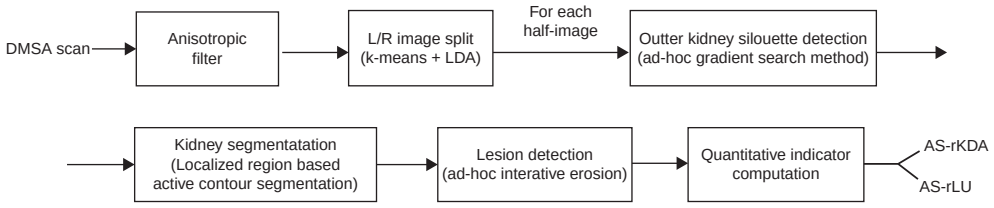


Fig. 2. Block diagram of the automatic DMSA kidney lesion detection and segmentation system.

images, a pixel-based free hand drawing tool with basic contour filling routines was sufficient for the physicians to segment the 16 pathological areas from each patient's DMSA posterior projection. Due to the high kidney size and tracer uptake distribution pattern variability across the group, physicians also segmented the whole damaged kidney in order to obtain relative measures (as explained below). Note that the damaged kidney segmentation needs not to be highly accurate since it will be only used as a size and global uptake reference.

Automatic segmentation and quantification methodology

A fully automatic segmentation framework is proposed and implemented as follows (the pipeline is summarized in Fig. 2). First, an anisotropic filter (Perona & Malik with 15 iterations, $\Delta t = 1/7$, and $\kappa = 30$)[12] was applied to the original DMSA scan in order to reduce its noise level while maintaining the underlying anatomical morphology.

Second, since lesions can be located in both left and right kidneys, they need to be isolated and analyzed separately. For that, an ad-hoc method to bisect the image aiming to separate both kidneys was implemented, based on running a linear discriminant analysis (LDA)[13] to find the natural separating line of both kidneys, which were previously labeled loosely by a pixel-location k-means[14] ($k = 2$) clustering algorithm. Note that while in most cases a simpler method would work (e.g. detecting the vertical line with a local minimum of tracer uptake near the center of the image and splitting in it), this more general method would work in more complex acquisition scenarios such as slightly rotated kidneys that could appear mildly overlapped in the scan.

Now, for each half-image obtained, the damaged kidney area (if any) should be detected and segmented. For that, in the first place, the kidney contour should be obtained. Special care should be taken in choosing the appropriate algorithm for that task (particularly in damaged kidneys), since the most common approaches will not include within the kidney segmentation the damaged area due to its low tracer uptake, thereby impeding the subsequent lesion detection and also altering the real kidney size.

Due to that fact, an ad-hoc image gradient search method was implemented for this task: from an initial seed pixel, a breadth-first search method is applied, performing an iterative procedure to include those neighbor pixels below a given gradient threshold (geodesic path,[22,23] using a threshold of 0.11) in the solution mask. Finally, a canny edge detector[24] is applied on the resulting mask to obtain the outer kidney's contour. This method was able to obtain successfully the outer contour of the kidneys, holding in its interior any possible kidney damage. Let this whole kidney segmentation be called S1.

Next, S1 was used as an initiation mask of a Localized Region Based Active Contour Segmentation algorithm (Lankton and Tannenbaum,[15] using $\alpha = 0.2$, the Yezzi energy function, and a maximum number of iterations of 5000), which isolated the non-damaged region of the kidney (S2). Note that the key strategy to be able to detect and segment any possible damaged area relies on the robust comparison of S1 and S2. If, after performing an iterative binary erosion[16] (using a disk with a radius of 3pixels as structuring element) of S1 until obtaining a maximum overlap[17] with S2 (let this eroded S1 mask be called S3), S2 and S3 have significantly different contours, this would be probably related to the presence of a lesion.

Therefore, the detection of a round-shaped connected component (CC[18]) on the difference mask D = S3@@-S2 will be associated with a pathological region. Note however that, given the noise conditions and the variable performance of all the image-processing steps applied until the computation of D, the presence of several connected components in D is expected. In order to select the correct CC that is properly associated with the lesion and discard the possible others, several rules are applied regarding the CC shape and location (described in the supplementary material).

Indicators definition and computation

From the manually-segmented (MS) regions (lesion and kidney contour), the following two indicators of SKD were proposed and computed: the relative kidney's damaged area (MS-rKDA), defined as the ratio of the pathological area and the kidney's area where the lesion is located; and the relative lesion uptake, defined as the ratio of the median image intensity of the pathological area and the one of the underlying damaged kidney (MS-rLU). The motivation for the computation of these two indications relates to the need of quantifying both the extent of the lesion and its tracer uptake. Analogously, form the automatically-segmented (AS) regions (lesion and kidney contour), the AS-rKDA and AS-rLU indicators were obtained. Note that, by definition, the MS-rKDA and MS-rLU values of the control DMSA scans is zero, since its obtention procedure is expert-guided. In contrast, the corresponding AS-rKDA and AS-rLU need not to be zero, and in fact will be used to evaluate the diagnostic power of the automatic quantification framework.

Finally, as mentioned in Section "Introduction", the derivation of an efficient indicator from a baseline DMSA scan that could predict the chronic character of the observed lesion would be of great interest. For this task, we first note that, although it cannot be stated as a general rule, large polar hypoactive areas without deformity of the outlines and with indistinct margins will generally heal whereas marked localized deformity of the outlines will generally correspond to permanent sequelae[19,20] (see Fig. 1 for an example).

Therefore, we propose the Dilated Lesion Histogram Exponent (DLHE) indicator as a predictor of renal sequelae. This parameter is defined as the exponent of a fitted exponential curve in the histogram of the segmented lesion after a dilation operation [16](using a disk-shaped structuring element of radius = 3 pixels). The rationale behind the proposal is that sharp transitions of intensity from physiological uptake to hypoactive areas (i.e. the renal lesions) would tend to be associated to volume loss and therefore related to permanent damage.

Further details about all the described image-processing algorithms' parameters and its derivation are available in the supplementary material.

Please cite this article in press as: Sampedro F, et al. Computing quantitative indicators of structural renal damage in pediatric DMSA scans. Rev Esp Med Nucl Imagen Mol. 2016. http://dx.doi.org/10.1016/j.remn.2016.06.010

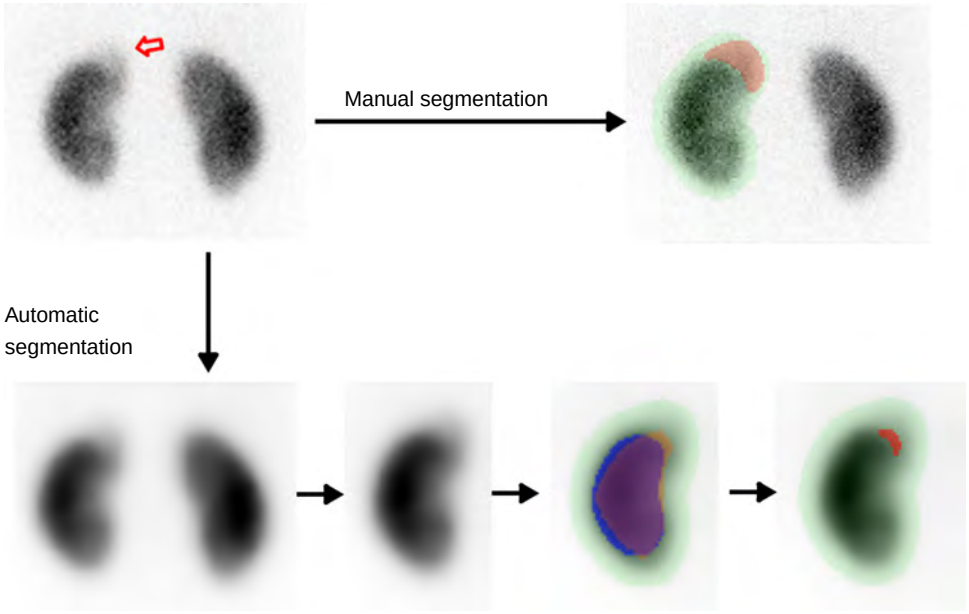


Fig. 3. DMSA scan (top left), its expert-guided manual segmentation (top right) and its automatic segmentation (down). Note the illustration of the aforementioned automatic segmentation steps: anisotropic filtering, image splitting, S1 (green)-S2 (blue)-S3 (orange) and the selected D's pathological CC (red).

Results

Segmentation results

Fig. 3 shows the manual and automatic segmentation results of a sample DMSA positive patient. Additional illustrative manual and automatic segmentation results for the DMSA scans of the study are available in the supplementary material.

Automatic segmentation accuracy with respect to the manual approach was highly variable across scans, given its aforementioned noise and morphological variability. Therefore, significant point-to-point correlation between MS-rKDA and AS-rKDA and MS-rLU and AS-rLU was not obtained ($p > 0.1$).

However, AS-rKDA achieved significant diagnostic power using a Receiver operating characteristic (ROC[21]) curve at the detection of positive DMSA scans, obtaining an Area Under the Curve (AUC) of 0.92 (Fig. 4). The cost-effective (cross point of sensitivity and specificity curves) AS-rKDA cut-off point derived from the ROC curve was 0.087, which obtained a sensitivity of 81% and a specificity of 94%.

Clinical correlations

We performed a correlation analysis between the proposed indicators (MS-rKDA, AS-rKDA, MS-rLU and AS-rLU) and the available clinical data. Two significant results were obtained. On the one hand, fever-negative patients had lower MS-rKDA than fever-positives ($p = 0.02$). On the other hand, a significant correlation ($p < 0.001$) was observed between the MS-rKDA and C-reactive protein (CRP) values of DMSA positive subjects (Fig. 5).

Chronic damage prediction

Fig. 6 illustrates the significant performance results of the DLHE indicator at detecting chronic lesions within the DMSA positive scans: AUC = 0.86, and sensitivity and specificity of 100% and 75%, respectively (computed from the cost-effective DLHE cutoff of 0.013). None of the other proposed indicators achieved significant performance within this context.

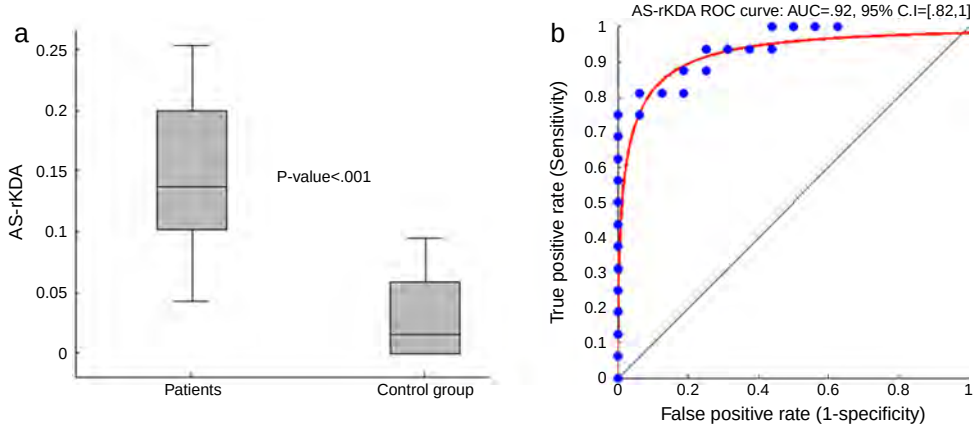


Fig. 4. (a) Whisker plot illustrating median AS-rKDA values, upper and lower quartiles, and minimum and maximum values for the patient's and control groups. (b) Receiver operating characteristic (ROC) AS-rKDA curve for the detection of DMSA positive scans. AUC: area under the curve; C.I.: confidence interval.

Please cite this article in press as: Sampedro F, et al. Computing quantitative indicators of structural renal damage in pediatric DMSA scans. Rev Esp Med Nucl Imagen Mol. 2016. http://dx.doi.org/10.1016/j.remn.2016.06.010

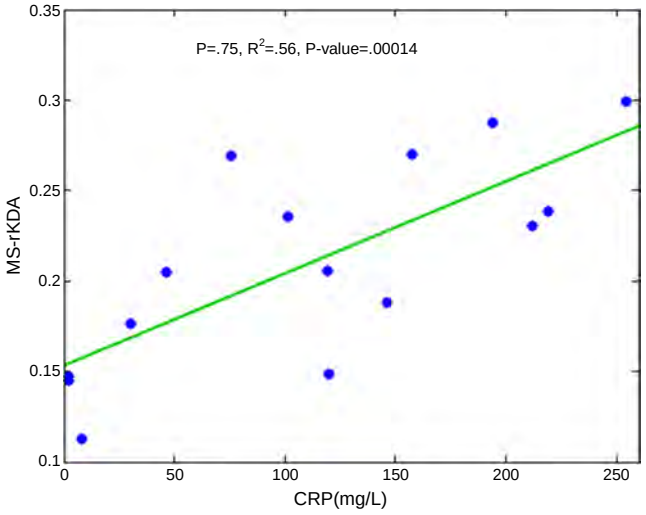


Fig. 5. Relation between MS-rKDA and CRP. ρ: Pearson's' correlation coefficient.

Discussion

In the present work we proposed, implemented and validated a DMSA image analysis framework for the computation of quantitative indicators that seek to characterize structural renal damage.

The set of proposed indicators was designed to model quantitatively the presence of a possible renal lesion in a DMSA scan. Considering both manual (MS-) and automatic (AS-) segmentation approaches, the relative kidney's damaged area (MS-rKDA, AS-rKDA) and its relative median intensity (MS-rLU, AS-rLU) were computed. Despite showing some accuracy limitations at the image-segmentation level, the automatically-computed AS-rKDA indicator was able to successfully classify most of the pathological and control DMSA scans, suggesting an innovation potential for a possible computed aided diagnosis tool in this particular scenario.

The biological significance of the MS-rKDA indicator was shown with its association with the CRP and fever values. The fact that the MS-rLU indicator did not correlate with any of the other non-imaging variables may suggest that this parameter, although having a conceptual basis, could be substantially distorted due to the DMSA acquisition variabilities. Hence, the relationship between the proposed DMSA-derived and additional nephrological parameters (such as renal ultrasound or histological patterns) needs to be further addressed in order to determine its potential clinical value.

Note that none of these indicators (MS-rKDA, AS-rKDA, MS-rLU, AS-rLU) succeeded in predicting the chronic character of the renal lesions. For that, a measure derived from the DMSA histogram of the segmented lesion (DLHE) was designed, which obtained a significant performance at modeling the underlying renal sequelae morphology. This result especially motivates further research in this line, given the potential advantages of the accurate chronic lesion identification from a single pediatric baseline DMSA scan.

This work has some limitations. On one hand, it used a relatively low number of subjects for the validation of the proposed image-quantification framework, which is especially critical in the

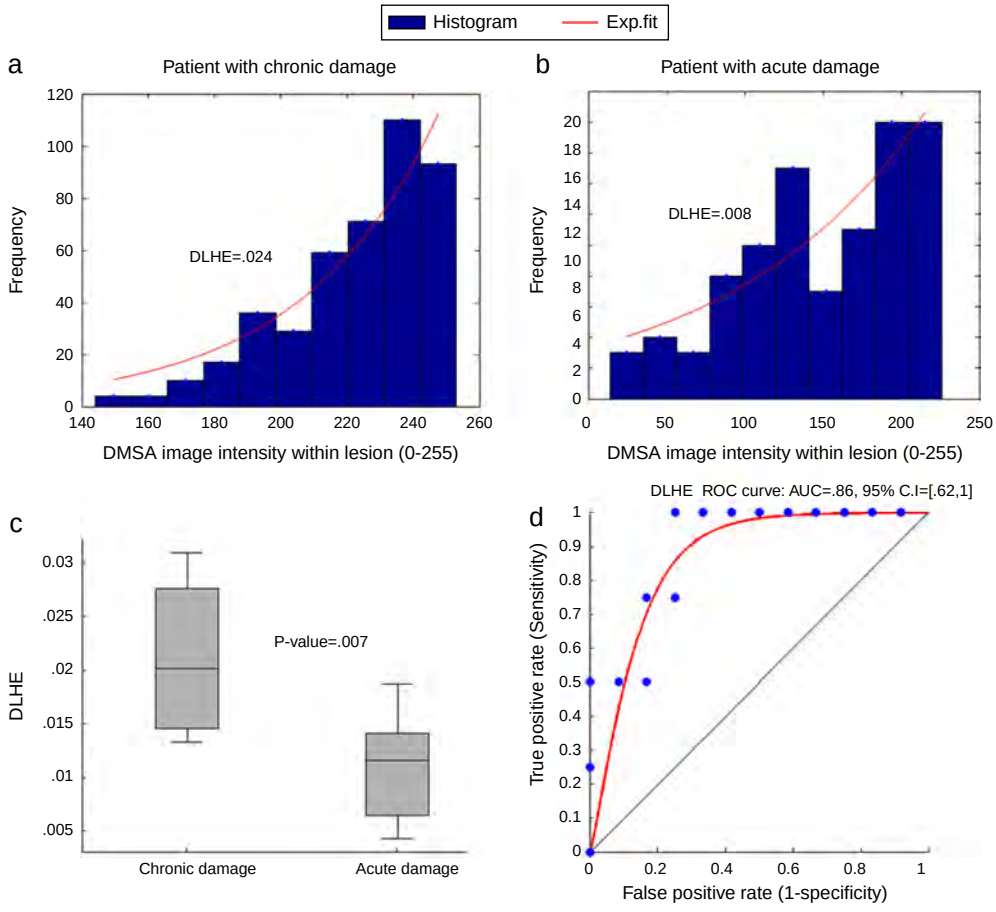


Fig. 6. (a and b) Dilated DMSA pathological area histograms and its exponential fit (DLHE being its exponent value) from an illustrative sample chronic and acute renal damaged patients. (c) Whisker plot illustrating median DLHE values, upper and lower quartiles, and minimum and maximum values for the chronic and acute damage groups. (d) ROC DLHE curve for the detection of chronic DMSA positive scans. AUC: area under the curve; C.I.: confidence interval.

Please cite this article in press as: Sampedro F, et al. Computing quantitative indicators of structural renal damage in pediatric DMSA scans. Rev Esp Med Nucl Imagen Mol. 2016. http://dx.doi.org/10.1016/j.remn.2016.06.010

assessment of the potential value of the DLHE parameter given the low number of patients with chronic lesions included in the study. Also, the fact that the visual DMSA evaluation was used as the gold-standard for the identification of structural renal damage and therefore used for the validation of the DMSA-derived quantitative indicators is another limitation of this study. In addition, only scans with clearly identifiable and upper-pole lesions were considered due to the computational challenge that represents the implementation of a generic DMSA lesion detection image-processing algorithm. On the other hand, a wider spectrum of nephrological variables should be used in the correlation analysis of the proposed indicator in order to fully appreciate their clinical value. To overcome these limitations, our proposals and results will need to be generalized and validated in large cohort (ideally multi-center), studies.

Taken together, this work proposed for the first time the implementation of a computational framework for the quantification of structural renal damage from DMSA scans. A set of image-derived numerical indicators was designed and computed on a group of pediatric DMSA scans, which showed a promising potential to complement visual DMSA evaluation and non-imaging renal damage indicators. Importantly, the incorporation of this type of image-quantification environment may provide major contributions on the early detection of permanent renal damage within the pediatric population.

Conclusion

In order to complement the visual diagnosis of structural renal damage from DMSA scans, a set of image-derived quantitative indicators was proposed. Its computation was performed within a novel computational framework that included both manual and automatic segmentation approaches. The performance of the proposed indicators at modeling the underlying renal pathology suggests a promising potential that will need to be validated and cross-validated in larger cohort studies.

Conflict of interests

The authors declare no conflict of interest.

Acknowledgements

The work of Frederic Sampedro is supported by the Spanish government FPU (Formación del Profesorado Universitario) doctoral grant (Grant No. AP2012-0400) and Spanish Project TIN2013-43478-P.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.remn.2016.06.010.

References

1. Saadeh SA, Mattoo TK. Managing urinary tract infections. Pediatr Nephrol. 2011;26:1967–76.
2. Sinha MD, Gibson P, Kane T, Lewis MA. Accuracy of ultrasonic detection of renal scarring in different centres using DMSA as the gold standard. Nephrol Dial Transplant. 2007;22:2213–6.
3. Czaja CA, Scholes D, Hooton TM, Stamm WE. Population-based epidemiologic analysis of acute pyelonephritis. Clin Infect Dis. 2007;45:273–80.
4. Uhari M, Nuutinen M. Epidemiology of symptomatic infections of the urinary tract in children. BMJ. 1988;297:450–2.
5. Bell LE, Mattoo TK. Update on childhood urinary tract infection and vesicoureteral reflux. Semin Nephrol. 2009;29:349–59.
6. Ditchfield MR, Summerville D, Grimwood K, Cook DJ, Powell HR, Sloane R, et al. Time course of transient cortical scintigraphic defects associated with acute pyelonephritis. Pediatr Radiol. 2002;32:849–52.
7. Gacinovic S, Buscombe J, Costa DC, Hilson A, Bomanji J, Ell PJ. Inter-observer agreement in the reporting of 99Tcm-DMSA renal studies. Nucl Med Commun. 1996;17:596–602.
8. Caglar M, Özgen Kıratlı P, Karabulut E. Inter- and intraobserver variability of ^{99m}Tc-DMSA renal scintigraphy: impact of oblique views. J Nucl Med Technol. 2007;35:96–9.
9. Aguiar P, Pérez-Fentes D, Garrido M, García C, Ruibal Á, Cortés J. A method for estimating DMSA SPECT renal function for assessing the effect of percutaneous nephrolithotripsy on the treated pole. Q J Nucl Med Mol Imaging. 2016;60:154–62.
10. Shaikh N, Ewing AL, Bhatnagar S, Hoberman A. Risk of renal scarring in children with a first urinary tract infection: a systematic review. Pediatrics. 2010;126:1084–91.
11. MATLAB 7.9 (R2009b) and Statistics Toolbox 7.2. Natick, MA, United States: The MathWorks, Inc.; 2009.
12. Perona P, Malik J. Scale-space and edge detection using anisotropic diffusion. IEEE Trans Pattern Anal Mach Intell. 1990;12:629–39.
13. Duda RO, Hart PE, Stork DH. Pattern classification. 2nd ed. Wiley Interscience; 2000.
14. Nameirakpam Dhanachandra, Khumanthem Manglem, Yambem Jina Chanu. Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. Procedia Comput Sci. 2015;54:764–71.
15. Lankton S, Tannenbaum A. Localizing region-based active contours. IEEE Trans Image Process. 2008;17.
16. Gonzalez RC, Woods RE, Eddins SL. Digital image processing using MATLAB. Gatesmark Publishing; 2009.
17. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index: scientific reports. Acad Radiol. 2004;11:178–89.
18. Haralick Robert M, Shapiro LG. Computer and robot vision, vol. 1. Addison-Wesley; 1992. p. 28–48.
19. Piepsz A, Colarinha P, Gordon I, Hahn K, Olivier P, Roca I, et al. Guidelines on 99mTc-DMSA scintigraphy in children. In: Paediatric Committee of the European Association of Nuclear Medicine. 2009.
20. Hitzel A, Liard A, Vera P, Manrique A, Menard JF, Dacher JN. Color and power Doppler sonography versus DMSA scintigraphy in acute pyelonephritis and in prediction of renal scarring. J Nucl Med. 2002;43:27–32.
21. Swets John A. Signal detection theory and ROC analysis in psychology and diagnostics: collected papers. Mahwah, NJ: Lawrence Erlbaum Associates; 1996.
22. Traumann A, Anbarjafari G, Escalera S. Accurate 3D measurement using optical depth information. Electron Lett. 2015.
23. Hernández A, Gatta C, Escalera S, Igual L, Martín-Yuste V, Radeva P. Accurate and robust fully-automatic QCA: method and numerical validation. In: MICCAI: 14th international conference on medical image computing and computer assisted intervention. 2011.
24. Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell. 1986;8:679–98.
25. Cao X, Zurakowski D, Diamond DA, Treves ST. Automatic measurement of renal volume in children using 99mTc dimercaptosuccinic acid SPECT: normal ranges with body weight. Clin Nucl Med. 2012;37:356–61.
26. Moskovitz B, Halachmi S, Sopov V, Burbara J, Horev N, Groshar D, et al. Effect of percutaneous nephrolithotripsy on renal function: assessment with quantitative SPECT of 99mTc-DMSA renal scintigraphy. J Endourol. 2006;20:102–6.

Please cite this article in press as: Sampedro F, et al. Computing quantitative indicators of structural renal damage in pediatric DMSA scans. Rev Esp Med Nucl Imagen Mol. 2016. http://dx.doi.org/10.1016/j.remn.2016.06.010

APOE-by-sex interactions on brain structure and metabolism in healthy elderly controls

Frederic Sampedro[1,2,3,*], Eduard Vilaplana[1,2,*], Mony J de Leon[4], Daniel Alcolea[1,2], Jordi Pegueroles[1,2], Victor Montal[1,2], María Carmona-Iragui[1,2], Isabel Sala[1,2], María-Belén Sánchez-Saudinos[1,2], Sofía Antón-Aguirre[1,2], Estrella Morenas-Rodríguez[1,2], Valle Camacho[3], Carles Falcón[5,7], Javier Pavía[6,7], Domènec Ros[5,7], Jordi Clarimón[1,2], Rafael Blesa[1,2], Alberto Lleó[1,2], Juan Fortea[1,2] for the Alzheimer's Disease Neuroimaging Initiative[**]

[1]Memory Unit, Department of Neurology, Hospital de la Santa Creu i Sant Pau- Biomedical Research Institute Sant Pau- Universitat Autònoma de Barcelona, Barcelona, Spain

[2]Centro de Investigación Biomédica en Red de Enfermedades Neurodegenerativas. CIBERNED, Madrid, Spain

[3]Nuclear Medicine Department, Hospital de la Santa Creu i Sant Pau- Biomedical Research Institute Sant Pau- Universitat Autònoma de Barcelona, Barcelona, Spain

[4]New York University School of Medicine, New York, NY, USA

[5]Unitat de Biofísica i Bioenginyeria, Departament de Ciències Fisiològiques I, Facultat de Medicina, Universitat de Barcelona – IDIBAPS, Barcelona, Spain

[6]Nuclear Medicine Department. Hospital Clínic de Barcelona, Barcelona, Spain

[7]Biomedical Research Networking Center in Bioengineering, Biomaterials and Nanomedicine - CIBER-BBN, Barcelona, Spain

[*]These authors have contributed equally to this work

[**]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Correspondence to:
Juan Fortea, **e-mail:** jfortea@santpau.cat

Keywords: *Gerotarget, Alzheimer's disease, aging, APOE, MRI, PET-FDG*

Received: *June 29, 2015* ***Accepted:*** *August 28, 2015* ***Published:*** *September 10, 2015*

ABSTRACT

Background: The *APOE* effect on Alzheimer Disease (AD) risk is stronger in women than in men but its mechanisms have not been established. We assessed the *APOE*-by-sex interaction on core CSF biomarkers, brain metabolism and structure in healthy elderly control individuals (HC).

Methods: Cross-sectional study. HC from the Alzheimer's Disease Neuroimaging Initiative with available CSF (*n* = 274) and/or 3T-MRI (*n* = 168) and/or a FDG-PET analyses (*n* = 328) were selected. CSF amyloid-ß$_{1-42}$ (Aß$_{1-42}$), total-tau (t-tau) and phospho-tau (p-tau$_{181p}$) levels were measured by Luminex assays. We analyzed the *APOE*-by-sex interaction on the CSF biomarkers in an analysis of covariance (ANCOVA). FDG uptake was analyzed by SPM8 and cortical thickness (CTh) was measured by FreeSurfer. FDG and CTh difference maps were derived from interaction and group analyses.

Results: *APOE4* carriers had lower CSF Aß$_{1-42}$ and higher CSF p-tau$_{181p}$ values than non-carriers, but there was no *APOE*-by-sex interaction on CSF biomarkers. The *APOE*-by-sex interaction on brain metabolism and brain structure was significant. Sex stratification showed that female *APOE4* carriers presented widespread brain hypometabolism and cortical thinning compared to female non-carriers whereas male

***APOE4* carriers showed only a small cluster of hypometabolism and regions of cortical thickening compared to male non-carriers.**

Conclusions: The impact of *APOE4* on brain metabolism and structure is modified by sex. Female *APOE4* carriers show greater hypometabolism and atrophy than male carriers. This *APOE*-by-sex interaction should be considered in clinical trials in preclinical AD where *APOE4* status is a selection criterion.

INTRODUCTION

The apolipoprotein E (*APOE*) genotype is the strongest genetic risk factor for Alzheimer's disease (AD) [1]. It has three isoforms, ε2, ε3 and ε4. The *APOE* ε4 allele (*APOE4*) increases the risk for AD [2]. The effect of the *APOE4* allele on AD biomarkers in healthy controls (HC) has been widely studied [3], [4]. *APOE4* carriers have consistently lower cerebrospinal fluid (CSF) ß-amyloid 1–42 ($Aß_{1–42}$) levels than non-carriers, but the differences in tau levels are more controversial [5]–[7]. Most, [8]–[10] but not all [18F]-fluorodeoxyglucose (FDG) PET studies [11]–[13] have shown hypometabolism in AD-related regions in *APOE4* carriers in late-middle age [8] and even earlier [10]. A gene-dosage effect on the hypometabolism has also been reported [9]. The relationship between the *APOE* genotype and brain structure is more controversial. Many cross-sectional studies have reported cortical thinning or hippocampal atrophy, [3], [4], [14] while several others have found no relationship [15] and two have reported increased gray matter in relation to the *APOE4* allele [16], [17].

Several factors might account for the conflicting results. First, the age-range differences between studies are critical because distinct effects of *APOE* across the lifespan have been described [18]. Not all brain changes associated with the *APOE* genotype reflect incipient AD. *APOE* has been implicated in normal human brain development [19]. Second, there are amyloid dependent [20] and independent [21] mechanisms underlying the *APOE* influences on AD risk. However, most studies assessing the role of *APOE* on brain structure and metabolism do not assess AD pathophysiological biomarkers to disentangle these mechanisms. Third, *APOE4* is likely to interact with other pathological factors, complicating the isolation of a unique genetic effect [4]. And fourth, some of the inconsistent imaging and biochemical findings related to *APOE* in HC might result from neglecting a possible *APOE*-by-sex interaction [6]. Most studies to date have included sex as a covariate in the analyses but they did not explicitly test for an *APOE*-by-sex interaction.

The finding that the *APOE* effect on AD risk is stronger in women than in men was reported in early studies, [22], [23] confirmed in meta-analyses, [23], [24] and in a recent longitudinal study [6]. However, only two studies have assessed *APOE*-by-sex interactions on AD biomarkers. Altmann et al found a significant interaction for tau in mild cognitive impairment patients [6]. Damoiseaux et al reported a significant *APOE*-by-sex interaction for CSF tau levels and default mode network abnormalities in healthy controls [25].

The interaction between *APOE4* and sex on brain structure and metabolism has not been established. This interaction could affect the design and interpretation of prevention trials in preclinical AD in which APOE is a selection criterion (i.e. the Alzheimer's Prevention Initiative *APOE4* Trial, NIH project number 1UF1AG046150–01). The aim of the present study was to examine the interactions between *APOE4* and sex on brain metabolism and structure, based on the hypothesis that the *APOE4* allele exerts a differential adverse effect on brain metabolism and structure depending on sex.

RESULTS

Demographic and clinical of the participants in the CSF, FDG and MRI subsets are summarized separately in the Table 1. CSF was available in 274 HC individuals, 328 had an FDG PET, 225 had a 3T MRI, and 137 subjects had all three biomarkers. There were no significant differences between the MRI, PET and CSF subsets in age, sex, *APOE* status, MMSE or CSF biomarkers. There were no significant differences in age, *APOE* status, MMSE or CSF biomarkers between males and females in all three subsets. In the FDG and CSF subsets, males had higher years of education than females ($p < 0.001$), but in the MRI subset this difference did not reach significance.

APOE4 carriers had lower CSF $Aß_{1–42}$ values than non-carriers in all three subsets ($p < 0.001$). *APOE4* carriers had higher CSF p-tau_{181p} values in the three subsets, but these only reached significance in the FDG and CSF subset which had larger sample sizes ($p < 0.001$ and $p = 0.004$ respectively). *APOE4* carriers had higher CSF t-tau values in the three subsets, but these only reached significance in the CSF subset ($p < 0.05$). There were no significant differences in MMSE scores or education between *APOE4* carriers compared to non-carriers in any of the subsets. There were no significant differences between males and females in CSF biomarkers. Neither was there an *APOE*-by-sex interaction on CSF $Aß_{1–42,}$ CSF t-tau or CSF p-tau_{181p} values in the analysis of covariance (ANCOVA) analyses.

Table 1: Demographic, cerebrospinal fluid and clinical data in the CSF, FDG-PET and MRI Alzheimer's Disease Neuroimage Initiative subsets.

		MRI (*N* = 168)	FDG-PET (*N* = 328)	CSF (*N* = 274)
***APOE4* N (%)**		50 (29.76%)	87 (26.5%)	71 (25.9%)
AGE		73.4 (6.02)	74.5 (5.57)	74.4 (5.97)
SEX (% Females)		53.6%	49.4%	50.4%
MMSE		29.1 (1.07)	29.0 (1.24)	29.1 (1.15)
YEARS OF EDUCATION		16.6 (2.55)	16.3 (2.77)	16.3 (2.69)
Aß$_{1–42}$* **	TOTAL	200.7 (49.92)	201.4 (52.46)	200.6 (52.51)
	APOE4−	211.3* (46.32)	213.5* (46.87)	212.1* (47.81)
	APOE4+	175.4* (49.58)	165.2* (51.85)	167.9* (51.87)
p-tau$_{p181}$* **	TOTAL	32.4 (16.41)	30.78 (18.14)	30.48 (17.97)
	APOE4−	31.3 (16.68)	28.3* (15.31)	28.2* (15.23)
	APOE4+	35.0 (15.62)	38.1* (23.38)	36.9*(23.10)
t-tau* **	TOTAL	66.0 (31.88)	68.9 (34.57)	68.4 (32.12)
	APOE4−	65.1 (32.60)	67.0 (34.84)	66.0** (30.29)
	APOE4+	68.2 (30.34)	74.5 (33.41)	75.1** (36.22)

APOE4+ = apolipoprotein E ε4 allele carrier, APOE− = apolipoprotein E ε4 allele non-carrier
Values are expressed as mean (standard deviation) unless specified.
*equals $p < 0.001$ and
**equals $p < 0.05$ for the *APOE4* carriers vs non-carriers comparison within each subset. Note that 137 subjects were included in the three subsets.
***CSF data only available in 146 subjects in the MRI subset and 242 subjects in the PET subset.

APOE-by-sex interaction on brain metabolism

Fig. 1A presents this FDG voxel-wise interaction analysis across the cerebral hemispheres, showing voxels with an *APOE*-by-sex interaction, covaried by age and years of education ($p < 0.005$, $k = 50$). Two clusters emerged, one located mainly in the anterior cingulate region and the other in the temporal region. To analyze the directionality, we isolated the temporal cluster, averaged the FDG uptake, and plotted it in box and whisker plots (Fig. 1B). As shown, this interaction was driven by the decreased metabolism in female *APOE4* carriers and the increased metabolism in male *APOE4* carriers. The main and interactive effects of *APOE4* status and sex on brain metabolism in the ANCOVA analysis were significant in the model (interaction term between *APOE4* status and sex: ß-coefficient = 0.069, standard error [SE] = 0.021, $p = 0.001$; main effect of *APOE4* status: ß-coefficient = –0.037, SE = 0.016, $p = 0.019$; main effect of sex: ß-coefficient = −0.041, SE = 0.018, $p = 0.026$). Similar results were found for the anterior cingulate cluster (not shown).

Fig. 2 shows the sex stratified *APOE4* group analyses for FDG, covaried by age and years of education. Female *APOE4* carriers showed widespread clusters of decreased metabolism ($p < 0.005$) across the whole cerebral cortex in both hemispheres with respect to *APOE4* non-carriers (Fig. 2A). Male *APOE4* carriers showed an isolated cluster of decreased metabolism ($p < 0.005$) in the precuneus with respect to non-carriers (Fig. 2B).

To examine the impact of CSF biomarkers in the *APOE*-by-sex interaction on brain metabolism, we included CSF Aß$_{1–42}$ and CSF p-tau$_{181p}$ as covariates in the analyses. The inclusion of the CSF biomarkers did not significantly alter the results of the *APOE*-by-sex interaction analysis (not shown) nor the female *APOE4* carriers vs non-carriers comparison (Fig. 3A1–3A3). In the male APOE4 carriers vs non-carriers comparison two clusters of increased metabolism emerged in *APOE4* carriers with respect to male non-carriers in prefrontal regions and a cluster in the medial temporal region when CSF Aß$_{1–42}$ levels or both Aß$_{1–42}$ and CSF p-tau$_{181p}$ levels (but not CSF p-tau$_{181p}$ levels alone, Fig. 3 B2) were included as a covariate (Fig. 3B1 and 3B3).

APOE-by-sex interaction on brain structure

Fig. 4A presents the vertex-wise interaction analysis across the whole cortical mantle, covaried by age and years of education, showing voxels with

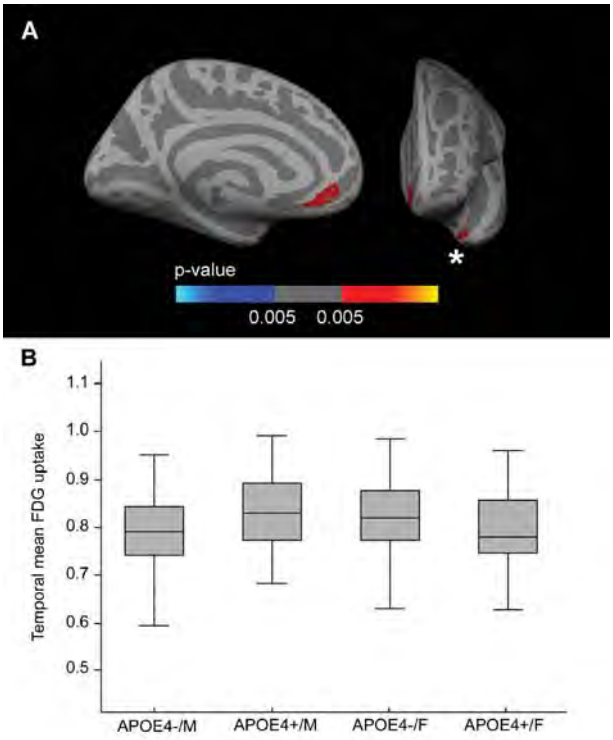


Figure 1: FDG *APOE*-by-sex interaction analysis. A. Areas in which there is a FDG-uptake interaction between sex and the *APOE4* status ($p < 0.005$ uncorrected) co-varied for age and years of education displayed across the medial and frontal views of the cerebral cortex. **B.** Box and whisker plot illustrating individual FDG-uptake values in the temporal cluster. For each plot, the central black lines show the median value, the regions above and below the black line show the upper and lower quartiles, respectively, and the whiskers extend to the minimum and maximum values. As illustrated, the female *APOE4* carriers showed decreased metabolism in the temporal cortex with respect to female non-carriers. FDG = fluorodeoxyglucose; *APOE* = apolipoprotein E, *APOE4+* = apolipoprotein E ε4 allele carriers, *APOE4−* = apolipoprotein E ε4 allele non-carriers.

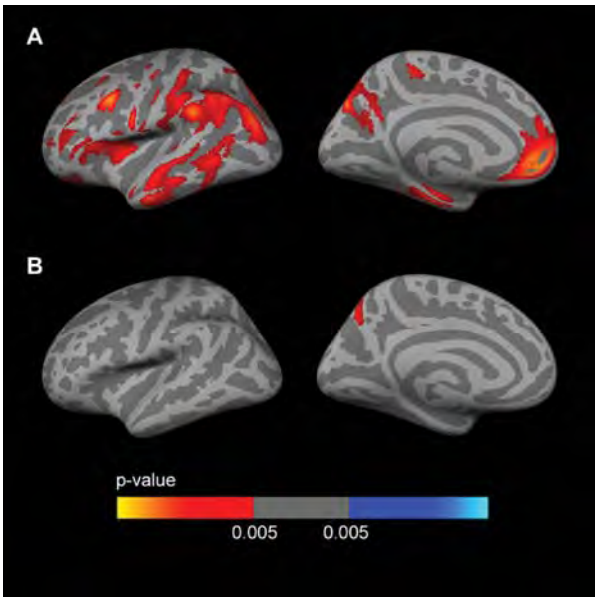


Figure 2: Sex-stratified FDG analyses. Analysis between *APOE4* carriers and *APOE4* non-carriers ($p < 0.005$ uncorrected) in **A.** females and **B.** males, co-varied for age and years of education across the lateral and medial views of the cerebral cortex. As shown, female *APOE4* carriers showed widespread clusters of decreased metabolism with respect to female *APOE4* non-carriers (Fig. 2A), whereas male *APOE4* carriers only showed an isolated cluster of decreased metabolism ($p < 0.005$) in the precuneus with respect to male non-carriers (Fig. 2B). FDG = fluorodeoxyglucose; *APOE4* = apolipoprotein E ε4 allele.

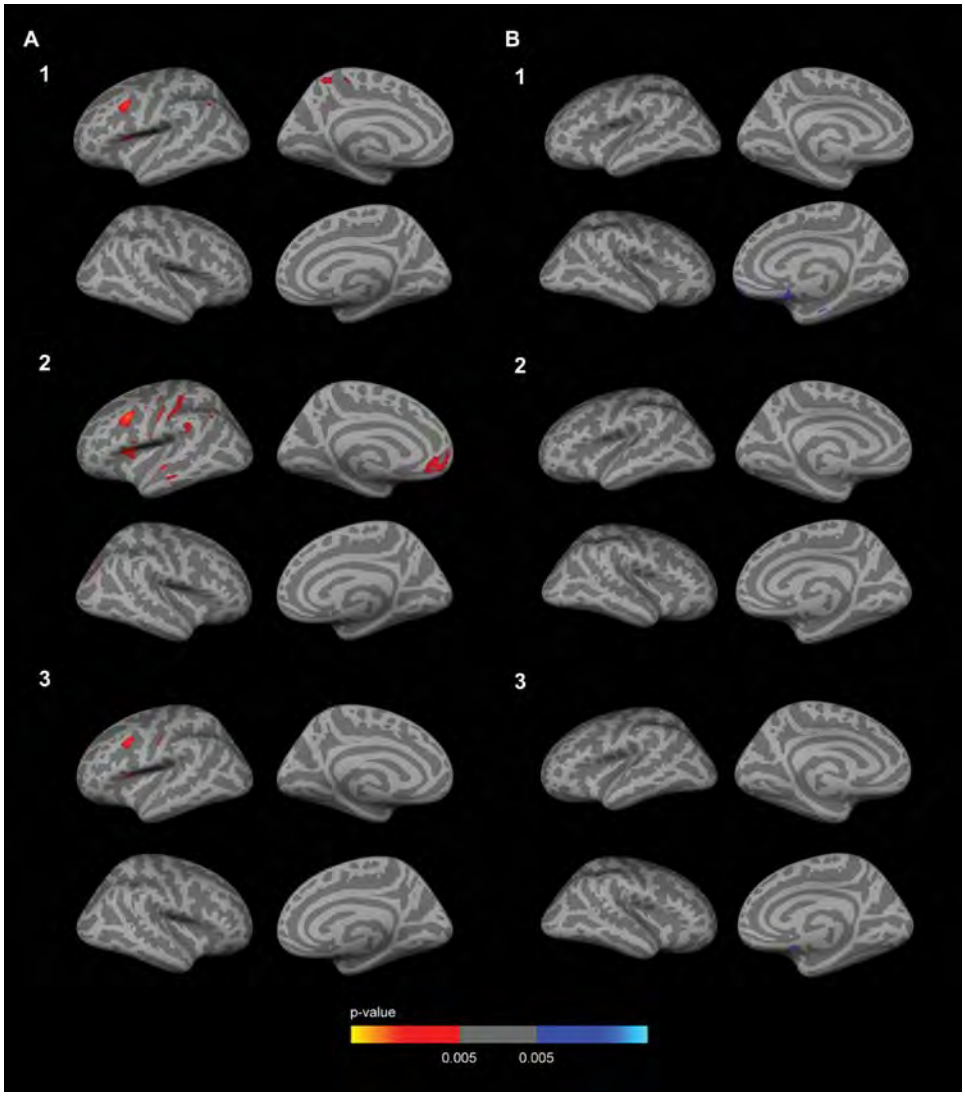


Figure 3: Sex-stratified FDG analyses with CSF biomarker levels included as a covariate. Row 1. CSF $A\beta_{1-42}$ levels; Row 2. CSF p-tau_{181p} levels; Row 3 CSF $A\beta_{1-42}$ and p-tau_{181p} levels. The analysis between female *APOE4* carriers and female *APOE4* non-carriers **A1-A3.** showed several clusters of decreased metabolism ($p < 0.005$ uncorrected) co-varied for age. As illustrated, female *APOE4* carriers showed decreased metabolism in the anterior cingulate cortex with respect to female non-carriers after the inclusion of the CSF biomarkers as a covariate. The analysis between male *APOE4* carriers and male *APOE4* non-carriers **B1-B3.** showed several clusters of increased metabolism ($p < 0.005$ uncorrected) co-varied for age. As illustrated, male *APOE4* carriers showed increased metabolism in several clusters in the dorsolateral prefrontal cortex with respect to male *APOE4* non-carriers after the inclusion of CSF $A\beta_{1-42}$ levels or both CSF $A\beta_{1-42}$ and CSF p-tau_{181p} as a covariate (B1 and B3), but not after the inclusion of the CSF p-tau_{181p} levels alone (B2). FDG = fluorodeoxyglucose; *APOE* = apolipoprotein E, *APOE4*: apolipoprotein E ε4 allele

an *APOE*-by-sex interaction. Two large clusters (Family-wise error corrected [FWE] $p < 0.05$) emerged, one in the dorsolateral frontal region and one in the temporoparietal region. To analyze the directionality, we then isolated the temporoparietal cluster, averaged the cortical thickness (CTh), and plotted it in a box and whisker plot (Fig. 4B). As shown, this interaction was mainly driven by the increased CTh in male *APOE4* carriers. The main effects and the interactive effects of *APOE4* status and sex in the ANCOVA analysis were significant in the model (interaction term between *APOE4* status and sex: ß-coefficient = −0.228, SE = 0.045, $p < 0.001$; main effect of sex: ß-coefficient = 0.149, SE = 0.039, $p < 0.001$; main effect of *APOE4* status: ß-coefficient = 0.062, SE = 0.030, $p = 0.041$). Similar results were found for the remaining cluster (not shown).

Fig. 5 shows the sex-stratified *APOE4* CTh analyses, covaried by age and years of education. Male *APOE4* carriers showed 3 large clusters (FWE corrected) of increased CTh with respect to non-carriers. Two of the clusters were observed in the left hemisphere, one in the dorsolateral frontal region and another in the

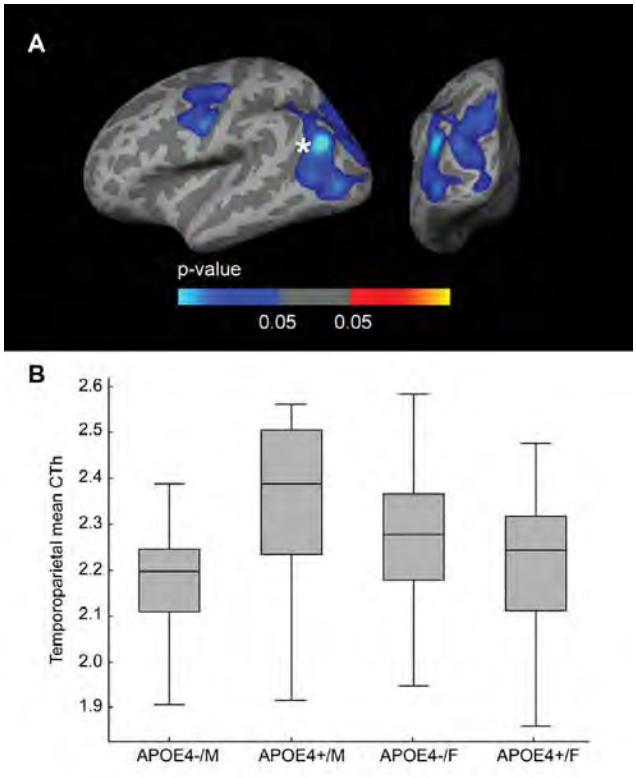


Figure 4: CTh *APOE*-by-Sex interaction analysis. A. Family-wise corrected ($p < 0.05$) clusters with an interaction between sex and the dichotomized *APOE4* genotype co-varied for age and years of education displayed across the lateral and posterior views of the cerebral cortex. **B.** Box and whisker plot illustrating individual CTh values in the temporo-parietal and occipital cluster. For each plot, the central black lines show the median value, regions above and below the black line show the upper and lower quartiles, respectively, and the whiskers extend to the minimum and maximum values. As illustrated, male *APOE4* carriers showed increased CTh in the temporo-parietal and occipital cluster. CTh = cortical thickness; *APOE* = apolipoprotein E, *APOE4+* = apolipoprotein E ε4 allele carriers, *APOE4−* = apolipoprotein E ε4 allele non-carriers.

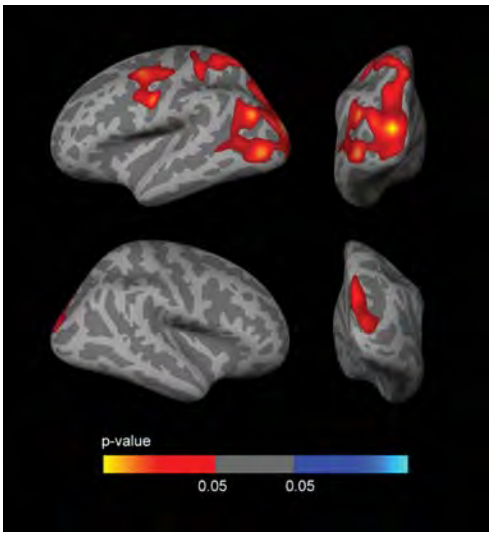


Figure 5: Sex-stratified CTh analyses. Analysis between male *APOE4* carriers and male *APOE4* non-carriers, co-varied for age and years of education. As shown, male *APOE4* carriers presented large clusters of increased CTh (FWE $p < 0.05$) in temporo-parieto-occipital regions, mainly in the left hemisphere. The analysis between female *APOE4* carriers and female *APOE4* non-carriers showed clusters of decreased CTh which did not survive FWE correction (not shown). CTh = cortical thickness; *APOE* = apolipoprotein E; FWE = family-wise error corrected ($p < 0.05$).

temporoparietal, occipital and precuneus regions. The third cluster was observed in the right hemisphere in the parietal and occipital regions. Female *APOE4* carriers showed cortical thinning in several regions than female *APOE4* non-carriers (not shown as this analysis did not survive FWE correction).

To examine the influence of CSF biomarkers on the *APOE*-by-sex interaction on brain structure, we included CSF $A\beta_{1-42}$ and CSF p-tau_{181p} as covariates in the analyses. The vertex-wise *APOE*-by-sex interaction analysis across the whole cortical mantle showed a reduction in the significance maps when including CSF biomarkers as covariates, especially $A\beta_{1-42}$ (Fig. 6). In the sex-stratified *APOE4* CTh analyses, the clusters of increased CTh in male *APOE4* carriers disappeared when CSF $A\beta_{1-42}$ levels (but not CSF p-tau_{181p} levels) were included as a covariate (Fig. 7). No result survived FWE correction in females.

All analyses were repeated excluding APOE ε2 allele carriers and including CSF t-tau as a covariate. We also restricted the analyses to non-hispanic white subjects (not shown). The results were not significantly altered in any case.

DISCUSSION

This study shows for the first time that the impact of the *APOE4* genotype on brain structure and metabolism is modified by sex. We found a significant *APOE*-by-sex interaction on brain metabolism and structure. Female *APOE4* carriers showed brain hypometabolism and cortical thinning with respect to female non-carriers whereas male *APOE4* carriers showed only a small cluster of hypometabolism and cortical thickening with respect to male non-carriers. CSF core AD biomarkers had an influence on brain structural results (and to a lesser extent on brain metabolism).

Epidemiologically, there is strong evidence that supports the *APOE*-by-sex interaction [6], [11], [23]. The only study assessing the *APOE*-by-sex interactions on MRI demonstrated the interaction on resting state functional connectivity but not on gray matter volume [25]. Our results expand these findings. We show an *APOE*-by-sex interaction on both brain structure and metabolism. The discrepancy on brain structure could be due to the differences in the subject population or technical differences (CTh analyses vs voxel-based morphometry [26]). Our FDG results are congruent with those of the aforementioned resting state functional connectivity analyses. *APOE* appears to affect brain network activity which is closely related to neuroenergetic functions [27].

Our metabolic findings suggest that women are metabolically more susceptible to the *APOE4* genotype. Neglecting a possible *APOE*-by-sex interaction on brain metabolism could be one of the reasons for the discordant FDG results [8]–[13]. Male *APOE4* carriers showed increased CTh and females decreased CTh. The finding of cortical thickening in AD vulnerable areas in middle aged (48–75 years old) *APOE4* carriers with respect to non-carriers has already been described [16], [17], but it is in contrast with other works assessing older cohorts [3], [4], [14], [15].

The discrepancies on brain structure might be conciliated if we consider a 2-phase phenomenon model in preclinical AD [28]. In this framework, pathological cortical thickening associated with low CSF $A\beta_{1-42}$ would be followed by atrophy once CSF p-tau_{181p} becomes abnormal [28]. Accordingly, our study shows that the clusters of increased CTh in male *APOE4* carriers disappear when we included CSF $A\beta_{1-42}$ as a covariate. The hypometabolism in female *APOE4* carriers did not disappear when CSF $A\beta_{1-42}$ levels were included as a covariate. The *APOE4* genotype might therefore exert its effects on brain glucose metabolism—at least in part—independently of amyloidogenic pathways [29]. Of note, the inclusion of CSF $A\beta_{1-42}$ levels as a covariate prompted the emergence of several areas of increased metabolism in male *APOE4* carriers. Increased brain metabolism in relation to brain amyloidosis has been previously described [30].

Altogether, our findings support that the mechanisms underlying the increased AD risk in female *APOE4* carriers might occur downstream of Aß pathology [6]. The *APOE4* effect on lowering CSF $A\beta_{1-42}$ levels is marked in both men and women (with no sex differences) and was also found in our work [6], [25]. The impact of an *APOE*-by-sex interaction on CSF has only been assessed twice and, as in the present work, always with data from the ADNI study. The absence of an *APOE*-by-sex interaction on CSF $A\beta_{1-42}$ levels is in agreement with the two previous works [6], [25]. The impact on CSF p-tau_{181p} levels is less clear. We did not find an *APOE*-by-sex interaction on CSF p-tau_{181p} levels. Such an interaction was reported initially [25] in HC but was not confirmed in the later work with a larger sample size [6]. Nonetheless, this last work did find the interaction for CSF p-tau_{181p} levels in mild cognitive impairment patients. Women, moreover, would be more susceptible and would present more abnormal neuronal injury biomarkers [25] and faster clinical decline [6]. Accordingly, female *APOE4* carriers showed hypometabolism and cortical thinning with respect to non-carriers, suggesting that female *APOE4* carriers might be more advanced in the aforementioned 2-phase phenomenon model in preclinical AD [28].

The mechanisms by which the *APOE* allele modifies the risk for AD have been extensively studied but are not completely understood. Both ß-amyloid-dependent [20] and ß-amyloid-independent [21] mechanisms have been described. *APOE* appears to affect brain network activity and neuroenergetic functions [27] and to increase microglia reactivity at Aβ plaques in mouse

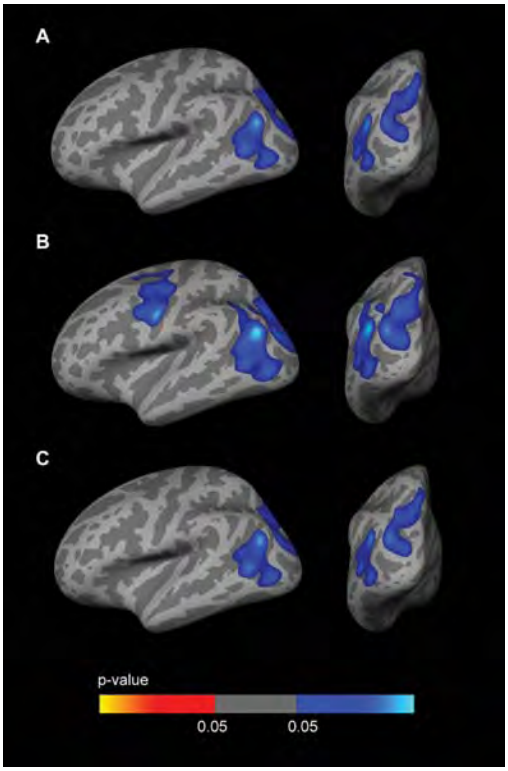


Figure 6: CTh *APOE*-by-Sex interaction analysis with CSF biomarker levels included as covariates. Family-wise corrected ($p < 0.05$) clusters with an interaction between sex and the dichotomized *APOE4* genotype co-varied for age and: **A.** CSF $A\beta_{1-42}$ levels; **B.** CSF p-tau_{181p} levels; **C.** CSF $A\beta_{1-42}$ and p-tau_{181p} levels. As illustrated, the inclusion of CSF $A\beta_{1-42}$ levels as a covariate significantly diminished the clusters showing a CTh *APOE*-by-sex interaction. CTh = cortical thickness; *APOE* = apolipoprotein E.

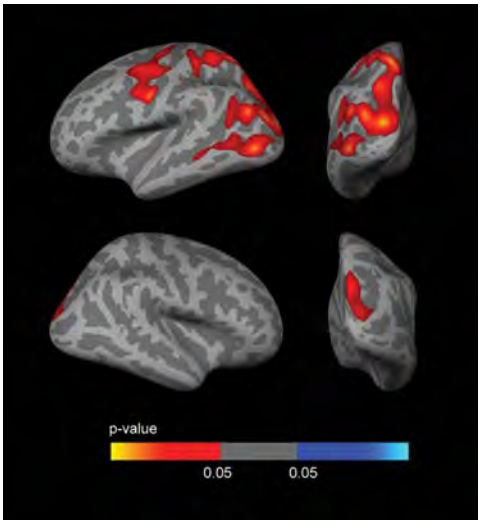


Figure 7: Sex stratified CTh analyses with CSF biomarker levels included as a covariate. The analysis between male *APOE4* carriers and male *APOE4* non-carriers showed several clusters of increased CTh ($p < 0.005$ uncorrected) co-varied for age and CSF p-tau_{181p} levels. There were no significant clusters of increased CTh male *APOE4* carriers vs male *APOE4* non-carriers after the inclusion of CSF $A\beta_{1-42}$ levels as a covariate. CTh = cortical thickness; *APOE* = apolipoprotein E.

models [31], [32]. These metabolic and inflammatory responses in relation to the *APOE* genotype might differ in males and females, accounting for the differences found.

This work has potential clinical implications. Clinical trials in preclinical AD in which *APOE4* status is a selection criterion are underway (Alzheimer's Prevention Initiative *APOE4* Trial, NIH project number 1UF1AG046150–01). Our results emphasize the importance of sex stratification when considering the AD risk and its impact on AD topographical biomarkers [33] conferred by the *APOE* genotype. More broadly, the present work stresses the need to consider interactions between biomarkers and risk factors in the AD preclinical phase [28].

The strengths of this study are the inclusion of a relatively high number of subjects and the fact that the results were found in two different topographical AD biomarkers, [34] with congruent findings between the two. The study has some limitations. It is cross-sectional and the age-range sampled does not include young HC to assess the age-range in which amyloid is starting to deposit in the brain of *APOE4* carriers [35].

In conclusion, the impact of *APOE4* on brain structure and metabolism is modified by sex in HC. This interaction should be considered in current clinical trials in preclinical AD in which *APOE4* status is a selection criterion.

MATERIALS AND METHODS

Study participants and clinical classification

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California – San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals (HC), people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see http://www.adni-info.org.

We included all HC with available CSF and/or a 3T-MRI and/or an FDG PET.

CSF analyses

ADNI procedure

Methods for CSF acquisition and biomarker measurement using the ADNI cohort have been reported previously [36]. $A\beta_{1-42}$, total tau (t-tau) and phospho-tau ($p\text{-}tau_{181p}$) levels were measured using the multiplex xMAP Luminex platform (Luminex) with Innogenetics (INNO-BIA AlzBio3) immunoassay kit–based reagents.

MRI and FDG-PET imaging procedures

ADNI acquisition procedure

The details of MRI and FDG-PET acquisition are available elsewhere (http://www.adni-info.org).

FDG-PET processing procedure

FDG-PET images were downloaded in the most processed format. They were intensity-scaled by the reference pons-vermis region [37], spatially normalized using SPM8 [http://www.fil.ion.ucl.ac.uk/spm/] to the Montreal Neurological Institute (MNI) PET template and spatially smoothed with a Gaussian kernel of full width at half-maximum (FWHM) of 8 mm. All resulting images were visually inspected to check for possible registration errors. Voxel-wise results were displayed at $p < 0.005$ (uncorrected) using an extent threshold $k = 50$, and projected on an inflated single-subject cortical surface reconstruction.

Cortical thickness processing procedure

Cortical reconstruction of the structural images was performed with the FreeSurfer software package, version 5.1 (http://surfer.nmr.mgh.harvard.edu). The procedures have been fully described elsewhere [38]. Estimated surfaces were inspected to detect errors in the automatic segmentation procedure. Fifty-seven of the 225 N3 processed MRI analyzed were excluded because

of segmentation errors and 168 were included in the analyses. A Gaussian kernel of 15 mm full-width at half maximum was applied. To avoid false positives, we tested Monte Carlo simulation with 10,000 repeats in Qdec (family-wise error [FWE], $p < 0.05$). Only regions that survived FWE are presented in the figures.

Statistical methods

Group analyses were made using SPSS (SPSS Inc, Chicago, IL). Comparisons between groups were performed using the two-tailed Student *t* test for continuous variables and a chi-square test for categorical variables.

The main objective of our work was to study the *APOE*-by-sex interaction on brain metabolism and brain structure. Two approaches were used: interaction and sex-stratified analyses. We carried out an ANCOVA as implemented in SPM and FreeSurfer for the PET and MRI analyses, respectively, using the *APOE* genotype (*APOE4* carrier vs *APOE4* non-carrier) and sex as binary categorical independent variables, and age and years of education as variables of no interest to assess the interaction.

To examine the impact of CSF biomarkers on the FDG PET and CTh analyses, we introduced CSF biomarkers as covariates in the analyses. All analyses were repeated excluding *APOE2* carriers and restricting to only non-hispanic white subjects.

Clusters derived from the interaction analyses in FDG or CTh were isolated to analyze the directionality of the interactive effects for each variable within an ANCOVA model, using age as a covariate. Specifically, we used the following model for FDG-PET and MRI:

Mean cluster FDG uptake (or mean cluster CTh) = $\hat{a}_0 + \hat{a}_1$*SEX + $\hat{a}_2$**APOE* + $\hat{a}_3$*[SEX**APOE*] + age

The same ANCOVA approach was used for the CSF analyses to test for an interactive effect of *APOE* genotype and sex in CSF biomarker levels.

ACKNOWLEDGMENTS

We thank Carolyn Newey for editorial assistance.

FUNDING

This work was supported by research grants from the Carlos III Institute of Health, Spain (grants PI11/02425 and PI14/01126 to Juan Fortea, grants PI10/1878 and PI13/01532 to Rafael Blesa and PI11/03035 to Alberto Lleó) and the CIBERNED program (Program 1, Alzheimer Disease to Alberto Lleó), partly funded by FEDER funds of the EU. This work has also been supported by a "Marató TV3" grant (531/U/2014 to Juan Fortea). The work of Frederic Sampedro is supported by the Spanish government FPU (Formación del Profesorado Universitario) doctoral grant (Grant No. AP2012–0400). This work was also supported by NIH-NIA grants to M.J. de Leon, AG022374, AG13616, and AG12101.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12–2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; andTakeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds tosupport ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

CONFLICTS OF INTEREST

All authors report no biomedical financial interests or potential conflicts of interest related to this work.

REFERENCES

1. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer,'s disease in late onset families. Science. 1993; Aug 261:921–3.
2. Liu C-C, Liu C-C, Kanekiyo T, Xu H, Bu G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat. Rev. Neurol. 2013; Feb 9:106–18.
3. Liu Y, Yu J-T, Wang H-F, Han P-R, Tan C-C, Wang C, Meng X-F, Risacher SL, Saykin AJ, Tan L. APOE genotype and neuroimaging markers of Alzheimer's disease: systematic review and meta-analysis. J. Neurol. Neurosurg. Psychiatry. 2015; May 86:127–34.

4. Reinvang I, Espeseth T, Westlye LT. APOE-related biomarker profiles in non-pathological aging and early phases of Alzheimer's disease. Neurosci. Biobehav. Rev. 2013; May 37:1322–1335.
5. Sunderland T, Mirza N, Putnam KT, Linker G, Bhupali D, Durham R, Soares H, Kimmel L, Friedman D, Bergeson J, Csako G, Levy JA, Bartko JJ, Cohen RM. Cerebrospinal fluid beta-amyloid1-42 and tau in control subjects at risk for Alzheimer's disease: the effect of APOE epsilon4 allele. BiolPsychiatry. 2004; Nov 56:670–6.
6. Altmann A, Tian L, Henderson VW, Greicius MD. Sex modifies the APOE-related risk of developing Alzheimer disease. AnnNeurol. 2014; Apr 75:563–73.
7. Vemuri P, Wiste HJ, Weigand SD, Knopman DS, Shaw LM, Trojanowski JQ, Aisen PS, Weiner M, Petersen RC, Jack CR. Effect of apolipoprotein E on biomarkers of amyloid load and neuronal pathology in Alzheimer disease. AnnNeurol. 2010; Mar 67:308–16.
8. Reiman EM, Caselli RJ, Yun LS, Chen K, Bandy D, Minoshima S, Thibodeau SN, Osborne . Preclinical evidence of Alzheimer's disease in persons homozygous for the epsilon 4 allele for apolipoprotein E. N. Engl. J. Med. 1996; Mar 334:752–8.
9. Reiman EM, Chen K, Alexander GE, Caselli RJ, Bandy D, Osborne D, Saunders AM, Hardy J. Correlations between apolipoprotein E epsilon4 gene dose and brain-imaging measurements of regional hypometabolism. Proc. Natl. Acad. Sci. U. S. A. 2005; Jun 102:8299–302.
10. Reiman EM, Chen K, Alexander GE, Caselli RJ, Bandy D, Osborne D, Saunders AM, Hardy J. Functional brain abnormalities in young adults at genetic risk for late-onset Alzheimer's dementia. Proc. Natl. Acad. Sci. U. S. A. 2004; Jan 101:284–9.
11. Corder EH, Jelic V, Basun H, Lannfelt L, Valind S, Winblad B, Nordberg A. No difference in cerebral glucose metabolism in patients with Alzheimer disease and differing apolipoprotein E genotypes. ArchNeurol. 1997; Mar 54:273–7.
12. Hirono N, Mori E, Yasuda M, Imamura T, Shimomura T, Hashimoto M, Tanimukai S, Kazui H, Yamashita H. Lack of effect of apolipoprotein E E4 allele on neuropsychiatric manifestations in Alzheimer's disease. J. Neuropsychiatry Clin. Neurosci. 1999; Jan 11:66–70.
13. Samuraki M, Matsunari I, Chen W-P, Shima K, Yanase D, Takeda N, Matsuda H, Yamada M. Glucose metabolism and gray-matter concentration in apolipoprotein E ε4 positive normal subjects. NeurobiolAging. 2012; Oct 33:2321–3.
14. Cherbuin N, Leach LS, Christensen H, Anstey KJ. Neuroimaging and APOE genotype: a systematic qualitative review. Dement. Geriatr. Cogn. Disord. 2007; Jan 24:348–62.
15. Novak NM, Stein JL, Medland SE, Hibar DP, Thompson PM, Toga AW. EnigmaVis: online interactive visualization of genome-wide association studies of the Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) consortium. Twin Res. Hum. Genet. 2012; Jun 15:414–8.
16. Espeseth T, Westlye LT, Fjell AM, Walhovd KB, Rootwelt H, Reinvang I. Accelerated age-related cortical thinning in healthy carriers of apolipoprotein E epsilon 4. NeurobiolAging. 2008; Mar 29:329–340.
17. Espeseth T, Westlye LTL, Walhovd KKB, Fjell AM, Endestad T, Rootwelt H, Reinvang I. Apolipoprotein E ε4-related thickening of the cerebral cortex modulates selective attention. NeurobiolAging. 2012; Mar 33:304–322.e1.
18. Filippini N, Ebmeier KP, MacIntosh BJ, Trachtenberg AJ, Frisoni GB, Wilcock GK, Beckmann CF, Smith SM, Matthews PM, Mackay CE. Differential effects of the APOE genotype on brain function across the lifespan. Neuroimage. 2011; Jan 54:602–10.
19. Dean DC, Jerskey BA, Chen K, Protas H, Thiyyagura P, Roontiva A, O'Muircheartaigh J, Dirks H, Waskiewicz N, Lehman K, Siniard AL, Turk MN, Hua X, Madsen SK, Thompson PM, Fleisher AS, Huentelman MJ, Deoni SCL, Reiman EM. Brain differences in infants at differential genetic risk for late-onset Alzheimer disease: a cross-sectional imaging study. JAMA Neurol. 2014; Jan 71:11–22.
20. Holtzman DM, Herz J, Bu G. Apolipoprotein E and apolipoprotein E receptors: normal biology and roles in Alzheimer disease. Cold Spring Harb. Perspect. Med. 2012; Mar 2:a006312.
21. Wolf AB, Valla J, Bu G, Kim J, LaDu MJ, Reiman EM, Caselli RJ. Apolipoprotein E as a β-amyloid-independent factor in Alzheimer's disease. Alzheimers. Res. Ther. 2013; Jan 5:38.
22. Poirier J, Davignon J, Bouthillier D, Kogan S, Bertrand P, Gauthier S. Apolipoprotein E polymorphism and Alzheimer's disease. Lancet. 1993; Sep 342:697–9.
23. Payami H, Montee KR, Kaye JA, Bird TD, Yu CE, Wijsman EM, Schellenberg GD. Alzheimer's disease, apolipoprotein E4, and gender. JAMA. 1994; May 271:1316–7.
24. Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, Pericak-Vance MA, Risch N, van Duijn CM. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. JAMA. 278:1349–561997.
25. Damoiseaux JS, Seeley WW, Zhou J, Shirer WR, Coppola G, Karydas A, Rosen HJ, Miller BL, Kramer JH, Greicius MD. Gender modulates the APOE ε4 effect in healthy older adults: convergent evidence from functional brain connectivity and spinal fluid tau levels. JNeurosci. 2012; Jun 32:8254–62.
26. Fortea J, Sala-Llonch R, Bartrés-Faz D, Bosch B, Lladó A, Bargalló N, Molinuevo JL, Sánchez-Valle R. Increased cortical thickness and caudate volume precede atrophy in PSEN1 mutation carriers. J. Alzheimers. Dis. 2010; Jan 22:909–22.

27. Wolf AB, Caselli RJ, Reiman EM, Valla J. APOE and neuroenergetics: an emerging paradigm in Alzheimer's disease. NeurobiolAging. 2013; Apr 34:1007–17.

28. Fortea J, Vilaplana E, Alcolea D, Carmona-Iragui M, Sánchez-Saudinos M-B, Sala I, Antón-Aguirre S, González S, Medrano S, Pegueroles J, Morenas E, Clarimón J, Blesa R, Lleó A. Cerebrospinal Fluid ß-Amyloid and Phospho-Tau Biomarker Interactions Affecting Brain Structure in Preclinical Alzheimer Disease. AnnNeurol. 2014; May :1–8.

29. Jagust WJ, Landau SM. Apolipoprotein E, not fibrillar β-amyloid, reduces cerebral glucose metabolism in normal aging. JNeurosci. 2012; Dec 32:18227–33.

30. Johnson SC, Christian BT, Okonkwo OC, Oh JM, Harding S, Xu G, Hillmer AT, Wooten DW, Murali D, Barnhart TE, Hall LT, Racine AM, Klunk WE, a Mathis C, Bendlin BB, Gallagher CL, Carlsson CM, a Rowley H, Hermann BP, Dowling NM, Asthana S, a Sager M. Amyloid burden and neural function in people at risk for Alzheimer's Disease. NeurobiolAging. 2014; Mar 35:576–84.

31. Rodriguez GA, Tai LM, LaDu MJ, Rebeck GW. Human APOE4 increases microglia reactivity at Aβ plaques in a mouse model of Aβ deposition. JNeuroinflammation. 2014; Jan 11:111.

32. Tai LM, Ghura S, Koster KP, Liakaite V, Maienschein-Cline M, Kanabar P, Collins N, Ben-Aissa M, Lei AZ, Bahroos N, Green S, Hendrickson B, Van Eldik LJ, LaDu MJ. APOE -modulated Aß-induced neuroinflammation in Alzheimer's disease: current landscape, novel data and future perspective. JNeurochem. 2015; Feb.

33. Sperling R, Aisen P, Beckett L. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups workgroups on diagnostic guidelines for Alzheimer's disease. 2011; May 7:280–92.

34. Dubois B, Feldman HH, Jacova C, Cummings JL, Dekosky ST, Barberger-Gateau P, Delacourte A, Frisoni G, Fox NC, Galasko D, Gauthier S, Hampel H, Jicha GA, Meguro K, O'Brien J, Pasquier F, Robert P, Rossor M, Salloway S, Sarazin M, de Souza LC, Stern Y, Visser PJ, Scheltens P. Revising the definition of Alzheimer's disease: a new lexicon. Lancet Neurol. 2010; Nov 9:1118–27.

35. Pletnikova O, Rudow G, Hyde T, Kleinman JE, Ali S, Bharadwaj R, Gangadeen S, Crain B, Fowler D, Rubio A, Troncoso J. Alzheimer lesions in the brains of young subjects. Cogn. Behav. Neurol. 2015; In press.

36. Shaw LM, Vanderstichele H, Knapik-Czajka M, Clark CM, Aisen PS, Petersen RC, Blennow K, Soares H, Simon A, Lewczuk P, Dean R, Siemers E, Potter W, Lee V M-Y, Trojanowski JQ, Initiative ADN. Cerebrospinal Fluid Biomarker Signature in Alzheimer's Disease Neuroimaging Initiative Subjects. AnnNeurol. 2009; Apr 65:403–413.

37. Landau S, Jagust W. UC Berkeley FDG MetaROI methods. Alzheimer's Dis. Neuroimaging Initiat. 2011.

38. Fischl B, Dale AM. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc. Natl. Acad. Sci. U. S. A. 2000; Sep 97:11050–5.

Brain Imaging and Behavior
DOI 10.1007/s11682-016-9607-5

ORIGINAL RESEARCH

Non-demented Parkinson's disease patients with apathy show decreased grey matter volume in key executive and reward-related nodes

Saul Martinez-Horta[1,2,3] · **Frederic Sampedro**[4] · **Javier Pagonabarraga**[1,2,3] · **Ramón Fernandez-Bobadilla**[1,2,3,5] · **Juan Marin-Lahoz**[1,2,3] · **Jordi Riba**[2,6] · **Jaime Kulisevsky**[1,2,3,4,5]

Abstract Apathy is a common but poorly understood neuropsychiatric disturbance in Parkinson's disease (PD). In a recent study using event-related brain potentials we demonstrated impaired reward processing and compromised mesocortico-limbic pathways in PD patients with clinical symptoms of apathy. Here we aimed to further investigate the involvement of reward circuits in apathetic PD patients by assessing potential differences in brain structure. Using structural magnetic resonance imaging (MRI) and voxel-based morphometry (VBM) we quantified grey matter volume (GMV) in a sample of 18 non-demented and non-depressed PD patients with apathy, and 18 matched non-apathetic patients. Both groups were equivalent in terms of sociodemographic characteristics, disease stage, cognitive performance and L-Dopa equivalent daily dose. Apathetic patients showed significant GMV loss in cortical and subcortical brain structures. Various clusters of cortical GMV decrease were found in the parietal, lateral prefrontal cortex, and orbitofrontal cortex (OFC). The second largest cluster of GMV loss was located in the left nucleus accumbens (NAcc), a subcortical structure that is a key node of the human reward circuit. Isolated apathy in our sample is explained by the combined GMV loss in regions involved in executive functions, and cortical and subcortical structures of the mesolimbic reward pathway. The correlations observed between apathy and cognition suggests apathy as a marker of more widespread brain degeneration even in a sample of non-demented PD patients.

Keywords Apathy · VBM · Parkinson's disease · Motivation · MRI · Behavior

Saul Martinez-Horta and Frederic Sampedro are contributed equally to this work

Electronic supplementary material The online version of this article (doi:10.1007/s11682-016-9607-5) contains supplementary material, which is available to authorized users.

✉ Jaime Kulisevsky
jkulisevsky@santpau.cat

1 Department of Neurology, Movement Disorders Unit, Hospital de la Santa Creu i Sant Pau Sant Antoni M. Claret 167, Universitat Autònoma de Barcelona, 08025 Barcelona, Spain

2 Sant Pau Institute of Biomedical Research (IIB-Sant Pau), Barcelona, Spain

3 Centro Investigación Biomedica en Red-Enfermedades Neurodegenerativas (CIBERNED), Barcelona, Spain

4 Faculty of Medicine, Autonomous University of Barcelona, Barcelona, Spain

5 Universitat Oberta de Catalunya (UOC), Barcelona, Spain

6 Human Neuropsychopharmacology Group IIB-Sant Pau, Barcelona, Spain

Introduction

Among the whole spectrum of behavioral disturbances found in Parkinson's disease (PD), apathy represents one of the most commonly reported (Pedersen et al. 2010; Aarsland et al. 2009). Apathy is defined as a state of diminished goal-directed behavior, reduced interest for pleasurable activities and flattened affect. These disturbances cannot be attributed to a decreased level of consciousness, cognitive impairment or depression (Marin 1991; Levy and Dubois 2006; Pagonabarraga et al. 2015). The prevalence of apathy in PD ranges from 17 % to 70 %, having a profound impact on the patient's quality of life and increasing the burden of

Published online: 11 October 2016

caregivers. Moreover, apathy severity has been associated with executive dysfunction and with an increased risk for the development of dementia. However, the executive deficits associated with apathy do not fully explain the clinical correlates and underlying mechanisms of apathy in PD.

Better understanding on the different brain circuits that cause apathy in PD would help to provide more adequate treatment strategies for its management (Dujardin et al. 2007). Signs and symptoms of apathy recorded from clinical observation have been structured into four subdomains involving: a) executive dysfunction (decrease in cognitive interests); b) deficits in auto-activation (lack of self-initiated mental processes); c) emotional distress (negative affect); and d) deficits in reward processing (decreases response to positive reinforces).

These four subdomains have been associated with different neural substrates. Executive dysfunction is thought to involve the dorsolateral prefrontal cortex (DLPFC), the dorsal caudate and putamen and the anterior cingulate cortex (ACC). Alterations in this circuit would lead to decreased planning and cognitive inertia (Levy and Dubois 2006). Impaired auto-activation has been associated with deficits in the ventral tegmental area, territories in the dorsomedial prefrontral cortex (PFC) including the supplementary motor area and the ACC. Alterations at this level would lead to a decrease in self-initiated behavior. The emotional distress subdomain has been related with hyperactivity in the subgenual cingulate cortex, and hypometabolism of the PFC and dorsal ACC. These alterations have been related to negative emotions of depression such as sadness and hopelessness. Finally, deficits in reward processing would involve the mesocortico-limbic pathway that includes the ventral tegmental area, orbitofrontal cortex, and nucleus accumbens (NAcc). The NAcc is a key node of the reward circuit, with robust activation responses to positive reinforcers (Riba et al. 2008).

The dysexecutive basis for apathy in PD has been clearly explained by the massive disruption of the dorsal caudate reciprocal thalamo-cortical projections to DLPFC. However, it has been also shown that alterations in the mesocortico-limbic pathway play an important role in the development of apathy (Martinez-Horta et al. 2014). Decreased responsiveness at this level would underlie emotional flatness, decreased emotional resonance, and decreased response to positive and negative reinforcers, as can be clinically observed in apathetic PD patients.

In line with reward processing deficits, in a recent study using event-related brain potentials we demonstrated reduced sensitivity to monetary incentives in early-stage PD patients with apathy (Martinez-Horta et al. 2014). The study compared cognitively-preserved and non-depressed PD patients with clinical symptoms of apathy with matched non-apathetic PD patients. The study showed significant decreases in the amplitude of the feedback-related negativity or FRN, a neurophysiological correlate of incentive processing. These results strongly supported a compromised mesocortico-limbic pathway as a key process in the pathogenesis of apathy in PD.

In the present study, we aimed to investigate the presence of structural brain abnormalities in PD patients who have developed clinically relevant symptoms of apathy. Using magnetic resonance imaging (MRI) and voxel-based morphometry (VBM) we compared brain structure between two groups of matched PD patients with isolated apathy. According to standard criteria, all participants were classified as non-demented and non-depressed and only differed with regard to the presence or absence of apathy.

Methods

Patient recruitment

Thirty-six PD patients with isolated apathy were prospectively included in the study. The sample was recruited from outpatients regularly visiting the Movement Disorders Unit at Sant Pau Hospital The diagnosis of PD was established according to the Queens Square Brain Bank criteria (Daniel and Lees 1993).

The diagnosis of apathy was established by using a semi-structured clinical interview based on the standard diagnostic criteria for apathy (P. Robert et al. 2009). An initial screening for the presence of clinically relevant symptoms of apathy was conducted using item 4 of the UPDRS part I (Goetz et al. 2008). The item is scored on a five-point scale ranging from 0 to 4, with higher scores indicating more severe symptoms of apathy. A score of 2/3 was chosen as an adequate value to initially identify potential study participants, avoiding the inclusion of patients with extreme symptomatology associated with the minimal score of 1 or the maximum score of 4. The score of 2/3 has adequate sensitivity and specificity (Leentjens et al. 2008) and a recent study confirmed it value for detecting apathy in PD (Weintraut et al. 2016). The semi-structured interview was given to screened patients, and only those fulfilling the diagnostic criteria for apathy were included in the study (sees supplementary material).

Exclusion criteria were patients presenting clinically meaningful depression and/or anxiety, as assessed by a score ≥ 11 on the depression and/or anxiety items of the Hospital Anxiety and Depression Scale (HADS) (Mumford 1991). More comprehensive assessment for depressive symptoms was done through the administration of a semi-structured clinical interview based on the standard DSM-IV-R diagnostic criteria for depression and dystimia. Presence of motor fluctuations in response to L-dopa, or medium-to-advanced PD according to Hoehn and Yahr stages (H&Y > 2) (Hoehn and Yahr 1967) also constituted exclusion criteria. Patients with dementia were also excluded, as assessed by a score < 24 on the Mini-Mental State Examination (MMSE) (Folstein et al. 1975), and a score < 123 on the Dementia Rating Scale

(DRS) (Llebaria et al. 2008), which constitutes a level 1 recommended instrument from the Movement Disorders Society Task Force for the screening of dementia in PD (PDD) (Litvan et al. 2011; Dubois et al. 2007; Emre et al. 2007). Patients with focal abnormalities in neuroimaging studies, alterations in blood tests, non-compensated systemic disease (i.e., diabetes, hypertension) and patients taking psychopharmacological medications were also excluded.

Each patient, with his or her caregiver if appropriate, was interviewed regarding disease onset and medication history, including type of motor response to L-dopa. All study participants were taking L-dopa and dopaminergic agonists (DA). Current medications and dosages were calculated for L-dopa daily dose, DA equivalent L-dopa daily dose and total L-dopa daily dose (LED) (Tomlinson et al. 2010). Participants were required to have received stable doses of dopaminergic drugs for the last 12 weeks and to show a stable response to medications. Motor status and disease stage were assessed by experienced neurologists in movement disorders (JP & JK) using the Unified Parkinson's Disease Rating Scale (UPDRS).

Potential differences between groups in demographic, clinical, cognitive and behavioral characteristics were analyzed with independent two-tailed t-tests for continuous variables, Mann-Whitney test for ordinal data, and the χ^2 test for categorical variables. Associations between the demographic, clinical and cognitive variables were analyzed with Pearson's correlations. Significance was set at $p < 0.05$.

MRI acquisition

T1-weighted images were acquired on a Phillips 3 T Achieva in sagittal orientation (TR = 7.4 and TE = 3.4, matrix size =228 mm × 218 mm; flip angle =9°, FOV = 250x250x180, slice thickness = 1.1 mm, 300 slices, acquisition time = 4′55″, voxel size =0.98 × 0.98 × 0.6).

MRI data processing and statistics

Gray matter volume (GMV) analysis from T1-weighted images was carried out using voxel-based morphometry (VBM) analysis in SPM8. The preprocessing steps were as follows.

First, unified segmentation was applied to the structural T1-weighted images of each subject. During this segmentation step, affine regularization was performed applying the values for the ICBM space template for European brains. The resulting tissue probability maps (GM maps) were then normalized to a standard stereotactic space using the corresponding DARTEL transformations to achieve spatial normalization into Montreal Neurological Institute (MNI) space. All normalized GM images were further analyzed to identify regional differences in GMV (using "modulation" to compensate for the effect of spatial normalization). Finally, the normalized and modulated images were smoothed using an isotropic spatial filter (FHWN =8 mm) to reduce residual inter-individual variability.

The individual smoothed GMV images were entered into a voxel-wise second-level two-sample t-test between the apathetic and non-apathetic PD patient groups. Individual values of total intracranial volume (TIV) were extracted and included as a nuisance variable to correct for global differences in TIV and, since no one of the recorded clinical variables exhibited significant differences between groups, age and sex were included as covariates of no interest. Results showing $p < 0.005$ (uncorrected) (Lieberman and Cunningham 2009) and a minimum extent of 50 voxels were considered significant. For the clusters showing significant gray matter differences, a small-volume correction (SVC) was applied (Worsley et al. 1996). Specifically, results were small volume corrected for family-wise error (FWE $p < 0.05$) within a sphere of 15 mm of diameter around peak coordinates extracted from independent studies (van der Vegt et al. 2013; Reijnders et al. 2010).

GMV at the regions of interest (ROIs) extracted from the clusters obtained in the former voxel-wise analysis were computed from build-in SPM8 functions to perform further regression analysis with other clinical variables of interest.

Results

Socio-demographic and clinical matching

As shown in Table 1, groups were carefully matched for all clinical and socio-demographic variables. Only the presence of symptoms of apathy differentiated the two groups. Data in the table are expressed as means ± standard deviation (SD) for the continuous variables, as percentage for the categorical variables and as mean range for the ordinal variables.

As indicated in the table, the sample was clinically characterized by individuals in the early to middle stages of the disease (disease duration 7.5 ± 5.1 years; H&Y stage 1.8 ± 0.4). In both groups, total MMSE and DRS scores ranged above the proposed cut-off score for dementia. To address the presence of subtle signs of cognitive impairment, we applied the accepted MDS criteria for mild cognitive impairment associated to PD (PD-MCI). (Litvan et al. 2011) Thus, using the suggested cut-off score of total DRS score < 138, up to 64 % of patients accomplished criteria for PD-MCI. Based on these criteria, prevalence of PD-MCI was up to 77 % in the apathy group and 50 % in the non-apathy group. These percentages resulted in a non-significant trend of increased prevalence of PD-MCI on the apathy group ($\chi^2 = 3.1$; $p = 0.083$). This result is consistent with previous findings supporting more impaired cognitive performance in apathetic PD patients (Pluck and Brown 2002; Martinez-Horta et al. 2013; Santangelo et al. 2015).

Table 1 Clinical and sociodemographic data

	Non-Apathy Group	Apathy Group	*p*
n	18	18	
Gender (f/m)[a]	10/8	10/8	$\chi^2 = .631$
Age (years)	64.8 ± 10.6	68.8 ± 10.1	.262
Education (years)	8.3 ± 3.6	10.5 ± 5	.150
Disease duration (years)	7.5 ± 5.1	5.1 ± 3	.080
MMSE[b]	28.2 ± 2.1	28.5 ± 1.6	.632
DRS[c]	134.2 ± 5.1	136.3 ± 5.5	.232
HADS-A[d]	8.1 ± 4.2	8.7 ± 4	.718
HADS-D[e]	4.8 ± 2.7	5.8 ± 3	.346
UPDRS Apathy score[f]	0	2.5 ± .5	< .000
H&Y stage[g]	1.8 ± .4	1.8 ± .3	.391
UPDRS III[h]	18.3 ± 6.3	20.8 ± 8	.326
L-dopa daily dose	375.8 ± 319	362.8 ± 339	.906
DA equivalent dose[i]	190 ± 216	169 ± 218	.773
Total LED[j]	565.8 ± 386	531.8 ± 397	.941

[a] Gender represented as number of females (f) and males (m)

[b] Mini mental state examination

[c] Dementia rating scale

[d] Hospital anxiety and depression scale – Anxiety score

[e] Hospital anxiety and depression scale – Depression score

[f] Item 4 unified Parkinson's disease rating scale

[g] Hoehn and yahr stage

[h] Unified Parkinson's disease rating scale total motor score

[i] Dopamine agonists L-dopa equivalent daily dose

[j] Total L-dopa daily equivalent dose. Data presented as *mean* ± *SD*

Focusing on specific subdomains of the DRS, a slight significant decrease was found for conceptualization in the apathy group ($p = .04$), but no significant differences were found on memory, attention, initiation/perseveration or construction.

No relevant signs of anxiety or depression were evidenced on the HADS scores and subsequent clinical interviews.

Voxel-based morphometry and statistical results

As seen in Fig. 1, between group comparisons of the VBM analysis showed a significant reduction of GMV in a set of cortical and subcortical brain regions in the apathetic patient group.

Cortical GMV decreases were found in the left inferior parietal lobule (BA 40, SVC $p = 0.002$ FWE), left superior parietal gyrus (BA 7), left orbitofrontal cortex (BA 47), right postcentral gyrus (BA 3), right inferior frontal gyrus (BA 44, SVC $p = 0.033$ FWE) and right supplementary motor area (Table 2).

The only subcortical structure that showed statistically significant differences between groups was the left nucleus accumbens (NAcc, SVC $p = 0.041$ FWE), in the ventral part of the striatum (Fig. 2). Using less stringent criteria ($p < 0.01$, uncorrected) GMV decreases were found bilaterally (see supplementary data).

GMV at the identified ROIs showed a significant correlation with clinical variables. Decreased GMV in the NAcc significantly correlated with global cognitive performance as measured by the DRS total score ($r = .831$; $p < .001$). Focusing on the different DRS sub-scores, this relation appeared exclusively associated with performance on the memory domain ($r = .388$; $p = .01$). At the emotional level, a negative correlation between the left inferior orbital prefrontal cortex and depression HADS scores ($r = -.523$; $p = .003$) was found. At the cognitive level, significant correlations were found between GMV reduction in the right inferior frontal gyrus with lower DRS total score ($r = .513$; $p = 0.001$), and lower conceptualization ($r = .371$; $p = .02$), memory ($r = .417$; $p = .01$) and initiation/perseveration ($r = .331$; $p = .04$) item scores. Decreased GMV in the left superior parietal gyrus correlated also with lower score in the initiation/perseveration item of the DRS ($r = .436$; $p = .008$).

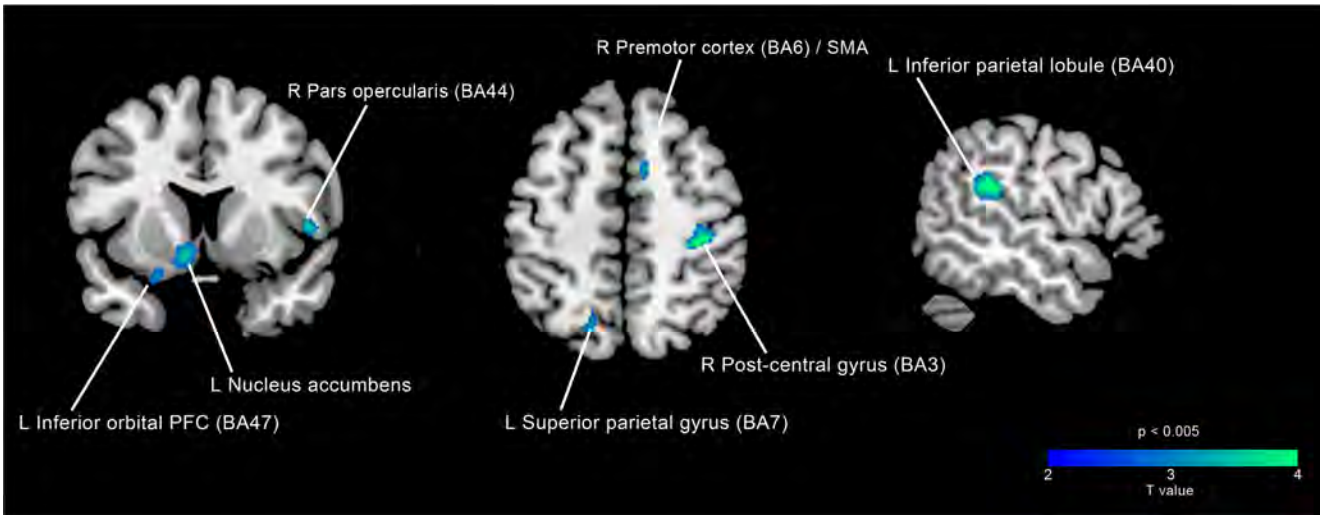


Fig. 1 Regions showing a reduction of gray matter volume in apathetic patients with respect to the non-apathetic group. There were no regions showing significant increase in gray matter volume

Discussion

In the present study, we searched for structural brain abnormalities in PD patients with clinical manifestations of isolated apathy. Based on previous data indicating deficits in reward processing in this population, we postulated that structural compromise will extend from territories linked to executive functions to structures within the mesocortico-limbic reward circuit. In keeping with this hypothesis, apathetic patients showed significant areas of GMV loss in subcortical and cortical brain regions. Significant clusters of GMV loss were located in the left NAcc and left inferior orbital PFC, both key nodes of the human reward circuit (Riba et al. 2008). Analysis on cortical findings showed spatially distributed clusters of grey matter decrease over the parietal and frontal lobes, involving functionally related areas that participate on action preparation/initiation, manipulation of information, as well as high-order integration of emotional stimuli.

Based on these findings, apathy in PD is associated with combined atrophy of fronto-parietal areas involved in executive functions, and regions of the human reward circuit. Initial research considered apathy in PD as a specific manifestation of the executive function caused by dopaminergic depletion of lateral prefrontal areas (Pluck and Brown 2002; Levy and Dubois 2006; Santangelo et al. 2015; Levy and Czernecki 2006). Apathy in PD has been consistently associated with the progressive executive dysfunction caused by decreased activation of lateral prefrontal and posterior parietal areas (Isella et al. 2002). In accordance, apathy in PD has been merely explained as secondary to functional deficits associated with nigrostriatal and mesocortical dopamine depletion in the putamen and caudate nucleus, respectively. (Santangelo et al. 2015; Pluck and Brown 2002; Martinez-Horta et al. 2013). However, when apathy in PD has been explored with more extensive neuropsychological batteries, it has also been observed to be associated with impairment in tasks involving reward or emotional processing (Martinez-Horta et al. 2013; Martinez-Corral et al. 2010). In agreement with this hypothesis, in recent study using event-related brain potentials we demonstrated reduced sensitivity to monetary incentives in apathetic PD patients. The study measured the amplitude of the feedback-related negativity (FRN) while participants

Table 2 Brain regions showing a significant reduction of grey-matter volume when comparing apathy vs. non-apathy groups

Brain area	Cluster size	Lateralization	BA	MNI (x, y, z)	Maximum t	*P* Value	FWE
Inferior parietal lobule	535	L	40	-57 -37 24	4.28	< 0.005	$p = 0.002$
Post-central gyrus	200	R	3	36–27 52	4.27	< 0.005	-
Pars opercularis	171	R	44	50 12 5	3.47	< 0.005	$p = 0.033$
Nucleus accumbens	412	L	-	-11 15–11	3.39	< 0.005	p = 0.041
Supplementary motor area	51	R	6	9 6 54	3.07	< 0.005	-
Inferior orbital PFC	93	L	47	-20 8–20	2.90	< 0.005	-
Superior parietal gyrus	91	L	7	-20 -70 42	2.85	< 0.005	-

BA Brodmann area, *MNI* Coordinates in montreal neurological institute stereotactic space. (p value <0.005; k = 50. FEW p value <0.05)

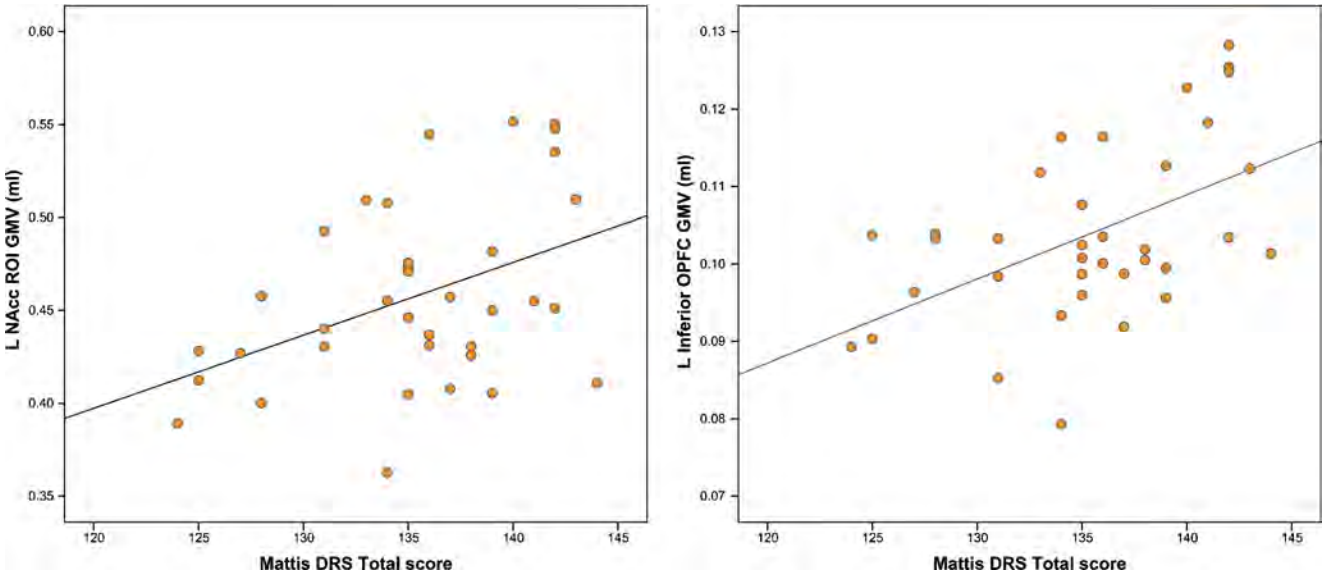


Fig. 2 Inverse correlation between cortical and subcortical GMV loss and cognitive performance

performed a lottery task. This wave, with generators in the ventral striatum and other limbic regions, was found to be significantly decreased in the apathetic subgroup (Martinez-Horta et al. 2014).

Our current anatomical findings give additional support to the notion of impaired reward processing in PD patients who develop apathy, and underline the existence of a more complex related circuitry which subserves motivational, cognitive and behavioral functions. The NAcc and the OFC are key structures within the mesolimbic reward pathway. In contrast with the nigrostriatal pathway, this circuit had been previously considered to remain relatively spared in early and middle PD stages (Rowe et al. 2008; Gotham et al. 1988, 1986). In contrast, the decreases in grey matter found here in both the NAcc and the OFC support its compromise in patients who develop apathy, even in the early stages of the disease. The left-sided lateralized pattern we found is consistent with the eminent unilateral-to-bilateral course of PD pathology. In fact, using less strict p-value ($p < 0.01$) a significant decrease can be bilaterally seen showing that right NAcc is not free of more severe degeneration in apathetic PD patients (see supplementary data).

These results are in line with recent neuroimaging studies using various assessment techniques. In one study using resting-state fMRI, Baggio and colleagues found an association between apathy and altered functional connectivity between the limbic regions of the PFC and the striatum (Baggio et al. 2015). However, this study did not properly control the effect of depression in the studied sample. In another study using shape analysis, the authors found atrophy of the NAcc in association with more severe apathetic symptoms in PD. However, part of the studied sample did not accomplish criteria for apathy, and this relationship was found only in relation to symptom severity (Carriere et al. 2014). Different PET studies have also given evidence on the decreased mesocortico-limbic dopaminergic activity present in apathetic PD patients. By using [11C]-Raclopride, [11C]-RTI-32 and [18]-FDG decreased dopamine release capacity has been observed in the mesolimbic circuit, as well as reduced binding and metabolism in the ventral striatum (Thobois et al. 2010; Remy et al. 2005; G. H. Robert et al. 2014). Our results extend and support the existence of structural abnormalities in the NAcc in non-demented PD patients from the early and middle stages of the disease.

In addition to the NAcc, we found GMV decreases in cortical brain areas. Atrophy in these regions may account for manifestations pertaining to other symptomatology domains than reward processing. Grey matter loss was found in the premotor cortex, including the SMA, and the *pars opercularis* of the inferior frontal gyrus (BA44). This cluster included regions around the insular cortex, the DLPFC and the *pars triangularis* (BA45). These last two areas connect with the middle (BA46) and the orbital (BA47) frontal areas. The alteration of the premotor cortex in our apathetic patients could be linked to the disruption of self-initiated behavior and thus to deficits in the auto-activation domain. These deficits would be further supported by cortical atrophy around the insula and related frontal structures. GMV decreases at this level would be consistent with difficulties in the executive integration of plans of action.

Cortical regions connected with the limbic system also showed loss of GMV. Within the medial prefrontal cortex, isolated apathy was associated with decreased GMV in the OFC. The OFC is part of the mesocortico-limbic reward circuit, playing a critical role in incentive processing and higher order integration of emotion (Timbie and Barbas 2015). Abnormalities in the OFC have been associated not only to apathy, but also to depression, anxiety, and social cognition

(Jenkins et al. 2014; Drevets 2007; Milad and Rauch 2007; Levy and Dubois 2006). The correlation observed between volume loss in the OFC and depression HADS scores could be interpreted as a marker of the emotional distress that may coexist even in apathetic patients without clinical criteria for depression(Pagonabarraga et al. 2015). On the contrary, it could be also the consequence that many items in commonly used scales for depression (including the HADS), are actually measuring decreased motivated behaviors. Since patients in our sample were free of clinically relevant depression, OFC atrophy –in conjunction with decreased volume in the NAcc– may indicate that not only loss of GMV in lateral aspects of the prefrontal cortex lead to apathy, but that the concurrent disruption of cortical and subcortical regions within the mesocortico-limbic reward are crucial for the clinical manifestation of decreased goal-directed behaviors.

Additional clusters of grey matter reduction were found in the inferior frontal gyrus and in the parietal lobes. These two structures have been associated with the cognitive aspects of apathy in PD (Pagonabarraga et al. 2015). Atrophy of these regions may account for previous evidence indicating a pattern of worse cognitive performance in apathetic patients (Martinez-Horta et al. 2013; Pluck and Brown 2002). These deficits were seen mainly in tasks involving frontal executive capacities, but also in others that rely on adequate parietal function (Martinez-Horta et al. 2013). Importantly, impairment in these tasks has been associated with more accelerated cognitive decline (Williams-Gray et al. 2009; Aarsland et al. 2011). This raises the question of a possible link between more severe global cognitive dysfunction and apathy. In PD various authors have shown that apathy may herald dementia (Williams-Gray et al. 2009; P. Robert et al. 2009), and grey matter atrophy and cortical thinning in posterior cortical regions have been associated with an increased risk of developing dementia (Aarsland et al. 2011; Bohnen et al. 2007; Bohnen et al. 2006). In the present study, the correlations observed between the NAcc and several cortical regions with global cognitive deterioration involving not only executive functions, suggests that the presence of apathy is a marker of more extensive cortical and subcortical degeneration even in a sample of non-demented patients.

Taken together, the present neuroimaging findings indicate the presence of structural abnormalities in PD patients with apathy. These abnormalities were observed in subcortical and cortical brain regions in a carefully selective sample of non-demented PD patients with isolated apathy in the early to mid-stages of the disease. GMV decreases in the NAcc demonstrate atrophy of a core structure of the mesocortico-limbic circuit and support a compromise of the reward circuit in this population. Areas of GMV decrease in the parietal lobe, as well as in the lateral and medial aspects of the prefrontal cortex fit well with the cognitive, auto-activation and emotional symptoms also present in apathy. Finally, the significant relation between structural changes and specific cognitive aspects links apathy to cognitive deterioration.

Given the highly specific characteristics of the patient subpopulation studied here, the present findings should be generalized with caution. Apathetic PD patients are only a subgroup of the broad range of PD patients usually encountered in the clinical practice. Also, symptom manifestations may evolve differently in the various domains that constitute apathy in the course of PD. Thus, the degree of compromise of the neural circuits discussed here may vary in the different stages of the disease.

Acknowledgments The work of Frederic Sampedro is supported by a Spanish Government FPU doctoral grant.

Compliance with ethical standards

Financial disclosures Saul Martinez-Horta, Frederic Sampedro, Ramon Fernandez-Bobadilla and Juan Marin-Lahoz declare that they have no conflict of interest.

Javier Pagonabarraga: Has received honoraria for lecturing or consultation from Boehringer Ingelheim, UCB, Allergan, Ipsen, and Lundbeck.

Jaime Kulisevsky: Has received honoraria for lecturing or consultation from the Michael J Fox Foundation, Merck Serono, AbbVie, Boehringer Ingelheim, UCB, Zambon, MSD, Italfarmaco, General Electric, and Lundbeck.

Informed consent All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, and the applicable revisions at the time of the investigation. Informed consent was obtained from all patients for being included in the study.

Conflict of interest All the authors report no conflict of interest.

Funding sources This study was partially funded by public research grants from CIBERNED (Fundación CIEN, Instituto de Salud Carlos III, Spain). The work of Frederic Sampedro is supported by a Spanish Government FPU doctoral grant.

References

Aarsland, D., Bronnick, K., Alves, G., Tysnes, O. B., Pedersen, K. F., Ehrt, U., et al. (2009). The spectrum of neuropsychiatric symptoms in patients with early untreated Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry, 80*(8), 928–930. doi:10.1136/jnnp.2008.166959.

Aarsland, D., Muniz, G., & Matthews, F. (2011). Nonlinear decline of mini-mental state examination in Parkinson's disease. *Movement Disorders, 26*(2), 334–337. doi:10.1002/mds.23416.

Baggio, H. C., Segura, B., Garrido-Millan, J. L., Marti, M. J., Compta, Y., Valldeoriola, F., et al. (2015). Resting-state frontostriatal functional connectivity in Parkinson's disease-related apathy. *Movement Disorders, 30*(5), 671–679. doi:10.1002/mds.26137.

Bohnen, N. I., Kaufer, D. I., Hendrickson, R., Ivanco, L. S., Lopresti, B. J., Constantine, G. M., et al. (2006). Cognitive correlates of cortical cholinergic denervation in Parkinson's disease and parkinsonian

dementia. *Journal of Neurology, 253*(2), 242–247. doi:10.1007/s00415-005-0971-0.

Bohnen, N. I., Kaufer, D. I., Hendrickson, R., Constantine, G. M., Mathis, C. A., & Moore, R. Y. (2007). Cortical cholinergic denervation is associated with depressive symptoms in Parkinson's disease and parkinsonian dementia. *Journal of Neurology, Neurosurgery, and Psychiatry, 78*(6), 641–643. doi:10.1136/jnnp.2006.100073.

Carriere, N., Besson, P., Dujardin, K., Duhamel, A., Defebvre, L., Delmaire, C., et al. (2014). Apathy in Parkinson's disease is associated with nucleus accumbens atrophy: a magnetic resonance imaging shape analysis. *Movement Disorders, 29*(7), 897–903. doi:10.1002/mds.25904.

Daniel, S. E., & Lees, A. J. (1993). Parkinson's Disease Society Brain Bank, London: overview and research. *Journal of Neural Transmission. Supplementum, 39*, 165–172.

Drevets, W. C. (2007). Orbitofrontal cortex function and structure in depression. *Annals of the New York Academy of Sciences, 1121*, 499–527. doi:10.1196/annals.1401.029.

Dubois, B., Burn, D., Goetz, C., Aarsland, D., Brown, R. G., Broe, G. A., et al. (2007). Diagnostic procedures for Parkinson's disease dementia: recommendations from the movement disorder society task force. *Movement Disorders, 22*(16), 2314–2324. doi:10.1002/mds.21844.

Dujardin, K., Sockeel, P., Devos, D., Delliaux, M., Krystkowiak, P., Destee, A., et al. (2007). Characteristics of apathy in Parkinson's disease. *Movement Disorders, 22*(6), 778–784. doi:10.1002/mds.21316.

Emre, M., Aarsland, D., Brown, R., Burn, D. J., Duyckaerts, C., Mizuno, Y., et al. (2007). Clinical diagnostic criteria for dementia associated with Parkinson's disease. *Movement Disorders, 22*(12), 1689–1707 quiz 1837. doi:10.1002/mds.21507.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*(3), 189–198.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., et al. (2008). Movement Disorder Society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement Disorders, 23*(15), 2129–2170. doi:10.1002/mds.22340.

Gotham, A. M., Brown, R. G., & Marsden, C. D. (1986). Levodopa treatment may benefit or impair "frontal" function in Parkinson's disease. *Lancet, 2*(8513), 970–971.

Gotham, A. M., Brown, R. G., & Marsden, C. D. (1988). 'Frontal' cognitive function in patients with Parkinson's disease 'on' and 'off' levodopa. *Brain, 111*(Pt 2), 299–321.

Hoehn, M. M., & Yahr, M. D. (1967). Parkinsonism: onset, progression and mortality. *Neurology, 17*(5), 427–442.

Isella, V., Melzi, P., Grimaldi, M., Iurlaro, S., Piolti, R., Ferrarese, C., et al. (2002). Clinical, neuropsychological, and morphometric correlates of apathy in Parkinson's disease. *Movement Disorders, 17*(2), 366–371.

Jenkins, L. M., Andrewes, D. G., Nicholas, C. L., Drummond, K. J., Moffat, B. A., Phal, P., et al. (2014). Social cognition in patients following surgery to the prefrontal cortex. *Psychiatry Research, 224*(3), 192–203. doi:10.1016/j.pscychresns.2014.08.007.

Leentjens, A. F., Dujardin, K., Marsh, L., Martinez-Martin, P., Richard, I. H., Starkstein, S. E., et al. (2008). Apathy and anhedonia rating scales in Parkinson's disease: critique and recommendations. *Movement Disorders, 23*(14), 2004–2014. doi:10.1002/mds.22229.

Levy, R., & Czernecki, V. (2006). Apathy and the basal ganglia. *Journal of Neurology, 253*(Suppl 7), VII54–VII61. doi:10.1007/s00415-006-7012-5.

Levy, R., & Dubois, B. (2006). Apathy and the functional anatomy of the prefrontal cortex-basal ganglia circuits. *Cerebral Cortex, 16*(7), 916–928. doi:10.1093/cercor/bhj043.

Lieberman, M. D., & Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social Cognitive and Affective Neuroscience, 4*(4), 423–428. doi:10.1093/scan/nsp052.

Litvan, I., Aarsland, D., Adler, C. H., Goldman, J. G., Kulisevsky, J., Mollenhauer, B., et al. (2011). MDS task force on mild cognitive impairment in Parkinson's disease: critical review of PD-MCI. *Movement Disorders, 26*(10), 1814–1824. doi:10.1002/mds.23823.

Llebaria, G., Pagonabarraga, J., Kulisevsky, J., Garcia-Sanchez, C., Pascual-Sedano, B., Gironell, A., et al. (2008). Cut-off score of the Mattis dementia rating scale for screening dementia in Parkinson's disease. *Movement Disorders, 23*(11), 1546–1550. doi:10.1002/mds.22173.

Marin, R. S. (1991). Apathy: a neuropsychiatric syndrome. *The Journal of Neuropsychiatry and Clinical Neurosciences, 3*(3), 243–254.

Martinez-Corral, M., Pagonabarraga, J., Llebaria, G., Pascual-Sedano, B., Garcia-Sanchez, C., Gironell, A., et al. (2010). Facial emotion recognition impairment in patients with Parkinson's disease and isolated apathy. *Parkinsons Disease, 2010*, 930627. doi:10.4061/2010/930627.

Martinez-Horta, S., Pagonabarraga, J., Fernandez de Bobadilla, R., Garcia-Sanchez, C., & Kulisevsky, J. (2013). Apathy in Parkinson's disease: more than just executive dysfunction. *Journal of the International Neuropsychological Society, 19*(5), 571–582. doi:10.1017/S1355617713000131.

Martinez-Horta, S., Riba, J., de Bobadilla, R. F., Pagonabarraga, J., Pascual-Sedano, B., Antonijoan, R. M., et al. (2014). Apathy in Parkinson's disease: neurophysiological evidence of impaired incentive processing. *The Journal of Neuroscience, 34*(17), 5918–5926. doi:10.1523/JNEUROSCI.0251-14.2014.

Milad, M. R., & Rauch, S. L. (2007). The role of the orbitofrontal cortex in anxiety disorders. *Annals of the New York Academy of Sciences, 1121*, 546–561. doi:10.1196/annals.1401.006.

Mumford, D. B. (1991). Hospital anxiety and depression scale. *The British Journal of Psychiatry, 159*, 729.

Pagonabarraga, J., Kulisevsky, J., Strafella, A. P., & Krack, P. (2015). Apathy in Parkinson's disease: clinical features, neural substrates, diagnosis, and treatment. *Lancet Neurology, 14*(5), 518–531. doi:10.1016/S1474-4422(15)00019-8.

Pedersen, K. F., Alves, G., Bronnick, K., Aarsland, D., Tysnes, O. B., & Larsen, J. P. (2010). Apathy in drug-naive patients with incident Parkinson's disease: the Norwegian ParkWest study. *Journal of Neurology, 257*(2), 217–223. doi:10.1007/s00415-009-5297-x.

Pluck, G. C., & Brown, R. G. (2002). Apathy in Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry, 73*(6), 636–642.

Reijnders, J. S., Scholtissen, B., Weber, W. E., Aalten, P., Verhey, F. R., & Leentjens, A. F. (2010). Neuroanatomical correlates of apathy in Parkinson's disease: A magnetic resonance imaging study using voxel-based morphometry. *Movement Disorders, 25*(14), 2318–2325. doi:10.1002/mds.23268.

Remy, P., Doder, M., Lees, A., Turjanski, N., & Brooks, D. (2005). Depression in Parkinson's disease: loss of dopamine and noradrenaline innervation in the limbic system. *Brain, 128*(Pt 6), 1314–1322. doi:10.1093/brain/awh445.

Riba, J., Kramer, U. M., Heldmann, M., Richter, S., & Munte, T. F. (2008). Dopamine agonist increases risk taking but blunts reward-related brain activity. *PloS One, 3*(6), e2479. doi:10.1371/journal.pone.0002479.

Robert, P., Onyike, C. U., Leentjens, A. F., Dujardin, K., Aalten, P., Starkstein, S., et al. (2009). Proposed diagnostic criteria for apathy in Alzheimer's disease and other neuropsychiatric disorders. *European Psychiatry, 24*(2), 98–104. doi:10.1016/j.eurpsy.2008.09.001.

Robert, G. H., Le Jeune, F., Lozachmeur, C., Drapier, S., Dondaine, T., Peron, J., et al. (2014). Preoperative factors of apathy in subthalamic stimulated Parkinson disease: a PET study. *Neurology, 83*(18), 1620–1626. doi:10.1212/WNL.0000000000000941.

Rowe, J. B., Hughes, L., Ghosh, B. C., Eckstein, D., Williams-Gray, C. H., Fallon, S., et al. (2008). Parkinson's disease and dopaminergic therapy–differential effects on movement, reward and cognition. *Brain, 131*(Pt 8), 2094–2105. doi:10.1093/brain/awn112.

Santangelo, G., Vitale, C., Trojano, L., Picillo, M., Moccia, M., Pisano, G., et al. (2015). Relationship between apathy and cognitive dysfunctions in de novo untreated Parkinson's disease: a prospective longitudinal study. *European Journal of Neurology, 22*(2), 253–260. doi:10.1111/ene.12467.

Thobois, S., Ardouin, C., Lhommee, E., Klinger, H., Lagrange, C., Xie, J., et al. (2010). Non-motor dopamine withdrawal syndrome after surgery for Parkinson's disease: predictors and underlying mesolimbic denervation. *Brain, 133*(Pt 4), 1111–1127. doi:10.1093/brain/awq032.

Timbie, C., & Barbas, H. (2015). Pathways for Emotions: Specializations in the Amygdalar, Mediodorsal Thalamic, and Posterior Orbitofrontal Network. *The Journal of Neuroscience, 35*(34), 11976–11987. doi:10.1523/JNEUROSCI.2157-15.2015.

Tomlinson, C. L., Stowe, R., Patel, S., Rick, C., Gray, R., & Clarke, C. E. (2010). Systematic review of levodopa dose equivalency reporting in Parkinson's disease. *Movement Disorders, 25*(15), 2649–2653. doi:10.1002/mds.23429.

van der Vegt, J. P., Hulme, O. J., Zittel, S., Madsen, K. H., Weiss, M. M., Buhmann, C., et al. (2013). Attenuated neural response to gamble outcomes in drug-naive patients with Parkinson's disease. *Brain, 136*(Pt 4), 1192–1203. doi:10.1093/brain/awt027.

Weintraut, R., Karadi, K., Lucza, T., Kovacs, M., Makkos, A., Janszky, J., et al. (2016). Lille apathy rating scale and MDS-UPDRS for screening apathy in Parkinson's disease. *Journal of Parasitic Diseases, 6*(1), 257–265. doi:10.3233/JPD-150726.

Williams-Gray, C. H., Evans, J. R., Goris, A., Foltynie, T., Ban, M., Robbins, T. W., et al. (2009). The distinct cognitive syndromes of Parkinson's disease: 5 year follow-up of the CamPaIGN cohort. *Brain, 132*(Pt 11), 2958–2969. doi:10.1093/brain/awp245.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping, 4*(1), 58–73. doi:10.1002/(SICI)1097-0193(1996)4:1<58::AID-HBM4>3.0.CO;2-O10.1002/(SICI)1097-0193(1996)4:1 < 58::AID-HBM4 > 3.0.CO;2-O.

Molecular Psychiatry (2015) 20, 772–777

www.nature.com/mp

OPEN

ORIGINAL ARTICLE

Telling true from false: cannabis users show increased susceptibility to false memories

J Riba[1,2,3], M Valle[2,3,4,11], F Sampedro[5,11], A Rodríguez-Pujadas[1], S Martínez-Horta[6], J Kulisevsky[6,7] and A Rodríguez-Fornells[8,9,10]

Previous studies on the neurocognitive impact of cannabis use have found working and declarative memory deficits that tend to normalize with abstinence. An unexplored aspect of cognitive function in chronic cannabis users is the ability to distinguish between veridical and illusory memories, a crucial aspect of reality monitoring that relies on adequate memory function and cognitive control. Using functional magnetic resonance imaging, we show that abstinent cannabis users have an increased susceptibility to false memories, failing to identify lure stimuli as events that never occurred. In addition to impaired performance, cannabis users display reduced activation in areas associated with memory processing within the lateral and medial temporal lobe (MTL), and in parietal and frontal brain regions involved in attention and performance monitoring. Furthermore, cannabis consumption was inversely correlated with MTL activity, suggesting that the drug is especially detrimental to the episodic aspects of memory. These findings indicate that cannabis users have an increased susceptibility to memory distortions even when abstinent and drug-free, suggesting a long-lasting compromise of memory and cognitive control mechanisms involved in reality monitoring.

Molecular Psychiatry (2015) **20,** 772–777; doi:10.1038/mp.2015.36; published online 31 March 2015

INTRODUCTION

Cannabis is the most widely used recreational drug worldwide after alcohol and tobacco.[1,2] Despite changing attitudes in the perceived risks associated with this substance and decriminalization initiatives taking place in many US states and countries,[1,3] the health implications of long-term cannabis consumption are still a matter of concern.[4] Regular use of cannabis has been associated with adverse health consequences, including psychiatric and neurocognitive disorders. Besides the more immediate risk of developing cannabis dependence,[5] other mental disorders, such as anxiety, depression or psychosis,[6,7] and cognitive impairment have also been described.[8] One recent study involving over a thousand individuals found that chronic cannabis use is associated with cognitive decline, with greater deterioration being observed in those individuals presenting a more persistent use.[9] Among the various cognitive domains studied, memory is one of the most frequently identified as being negatively affected by cannabis.[9–11]

Impaired working and declarative memory are well-known aspects of acute intoxication.[12] Cannabis preparations and delta-9-tetrahydrocannabinol, its main active principle, acutely deteriorate the ability to retain information for short periods of time,[8,13] and impair episodic memory and verbal recall.[14,15] A characteristic of cannabis consumption is that residual effects can linger for days after the most recent use.[10] Typically, these deleterious effects gradually wear off and memory processes normalize after several weeks of abstinence.[16,17] However, some studies in heavy cannabis users have observed impairment persisting even months after the last consumption.[9,10] In addition to impaired performance, imaging studies in chronic cannabis users have found structural brain alterations in the hippocampus, a key area in the memory processing network. Notably, decreases in hippocampal volume showed an association with the amount of cannabis used.[18–20] These structural changes may be long-lasting, as volume reductions can persist even after abstinence of 6 months.[18]

An unknown aspect of long-term cannabis use is its potential to disrupt memory and reality monitoring mechanisms that normally allow us to distinguish between veridical and illusory events. Avoiding memory distortions may be extremely relevant in certain contexts such as the courtroom and forensic examination, and in a more general context this ability provides us with an adequate sense of reality that guides future behavior based on past experiences. Memories of events that never occurred, or false memories, can be found in neurological and psychiatric conditions. They have been described in post-traumatic stress disorder, psychosis, dissociative disorders and in cases of confabulation or 'honest lying' associated with confessions of uncommitted crimes, among others.[21] However, in a more subtle form, false memories are also a common occurrence in everyday life in healthy individuals[22] and show an increase with age.[23] Susceptibility to this phenomenon probably has a neural basis, as it has been linked to individual differences in white matter microstructure.[24]

[1]Human Neuropsychopharmacology Group, Sant Pau Institute of Biomedical Research (IIB-Sant Pau), Sant Antoni María Claret 167, Barcelona, Spain; [2]Centre d'Investigació de Medicaments, Servei de Farmacologia Clínica, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain; [3]Departament de Farmacologia i Terapèutica, Universitat Autònoma de Barcelona, Centro de Investigación Biomédica en Red de Salud Mental, CIBERSAM, Barcelona, Spain; [4]Pharmacokinetic and Pharmacodynamic Modelling and Simulation, IIB-Sant Pau, Sant Antoni María Claret, Barcelona, Spain; [5]School of Medicine, Universitat Autònoma de Barcelona, Barcelona, Spain; [6]Movement Disorders Unit, Neurology Department, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona. Barcelona, Spain; [7]Centro Investigación Biomedica en Red-Enfermedades Neurodegenerativas (CIBERNED), Spain; [8]Cognition and Brain Plasticity Group (Bellvitge Biomedical Research Institute) IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain; [9]Department of Basic Psychology, University of Barcelona, Barcelona, Spain and [10]Catalan Institution for Research and Advanced Studies, ICREA, Barcelona, Spain. Correspondence: Dr J Riba, Human Neuropsychopharmacology Group, Sant Pau Institute of Biomedical Research (IIB-Sant Pau), IIB-Sant Pau. C/Sant Antoni María Claret, 167, Barcelona 08025, Spain.
E-mail: jriba@santpau.cat
[11]These authors contributed equally to this work.
Received 29 October 2014; revised 5 February 2015; accepted 13 February 2015; published online 31 March 2015

False memories can be induced in laboratory conditions using experimental procedures such as the Deese-Roediger-McDermott paradigm.[25] In this task, participants study a list of words that are later presented together with semantically unrelated new words and semantically related new words (lures).[25] Lures induce the illusion of a false memory where participants mistakenly claim that the new stimulus has been encountered previously. The correct identification of lures as previously unseen stimuli is more cognitively demanding than that of unrelated novel stimuli, the former leading to greater activation of medial temporal lobe (MTL), parietal and frontal brain regions.[26] In the present study we tested susceptibility to false memories in a group of abstinent heavy cannabis users and their matched controls using the Deese-Roediger-McDermott paradigm in association with functional magnetic resonance imaging (fMRI; see online methods).

MATERIALS AND METHODS

Ethics

The study was approved by the Ethics Committee of Sant Pau Hospital and all participants gave their written consent to participate.

Participants

We recruited a group of 16 heavy cannabis users not seeking or having a history of treatment for their cannabis consumption. We defined heavy cannabis use as daily use for at least the last 2 years. The recruited sample had never been diagnosed with a psychiatric or neurological condition including alcohol or other drug abuse. Cannabis users were matched to a cannabis-naive (<50 occasions of cannabis use in their lifetime) group of healthy controls, free of psychiatric or neurological conditions. Fourteen controls had used cannabis <10 times and only two had used it between 10 and 50 times. To rule out a history of psychiatric and neurological disorders, users and controls were interviewed by a clinical psychologist. The two groups were matched taking into account the following socio-demographic variables: sex, age, years of education, verbal intelligence and fluid intelligence. Verbal intelligence was assessed using a Spanish version of the NART,[27] known as TAP–'Test de Acentuación de Palabras' ('word accentuation test').[28] Fluid intelligence was assessed using a computerized version of the Matrix Reasoning from the Wechsler Adult Intelligence Scale-III.[29] Detailed socio-demographic data for each group is provided as Supplementary information.

Cannabis users had taken the drug an average of around 42 000 times (range: 4 000–246 375) times. The average number of years of use was 21 (3–39). The average number of daily cannabis cigarettes smoked was 5 (1–24) and the average age of initial use was 17 (12–20) years. We did not exclude tobacco smokers from the study and they were not instructed to abstain from tobacco during the study. Ten participants in the cannabis group and four in the control group were currently using tobacco. Participants abstained from cannabis use for at least 4 weeks prior to testing. Urine samples were taken during the 4-week period and immediately before the experimental session. All participants tested negative for cannabis, alcohol, benzodiazepines, amphetamines, opiates and cocaine on their day of participation.

Memory paradigm

The memory paradigm consisted in a modified version of the Deese-Roediger-McDermott paradigm[25] and included a study phase and a testing phase (see Supplementary information). Both phases were conducted with the participant in the MRI scanner. Stimuli were presented using goggles and behavioral responses were recorded by button press using a magnet-compatible response pad.

The study phase comprised 20 lists of four words. Prior to the presentation of the four words comprising a list, the name of that list was announced on the screen. Of the 20 lists, fifteen comprised four semantically related Spanish words and the other 5 lists comprised 3 semantically related words plus a catch word. Catch words were semantically unrelated to the list announced and were used to control for the participant's attention during the task. A total of 80 stimuli were presented during the study phase: 75 legitimate words plus 5 catch words. Participants were requested to indicate by button press whether the presented word belonged to the announced list. The order of presentation of the 20 word lists was randomized between participants. The study phase lasted 11 min.

Approximately 15 min after completion of the study phase, the test phase was conducted and lasted 14 min. Participants were presented with the 75 legitimate words shown during the study phase plus 40 semantically unrelated new words and 40 semantically related new words (lures, see stimuli tables in the Supplementary information file). Stimuli were presented in semi-random order with the restriction that the same type of stimulus (old, new or lure) was not presented more than twice in succession. We used a rapid presentation event-related design. Stimulus duration was 500 ms. The stimulus onset asynchrony was on average 5.125 s and it was jittered between 4 s and 10 s. The order and timing of events were optimized using the Optseq2 software (http://surfer.nmr.mgh.harvard.edu/optseq/). Participants were required to judge whether a word had been presented in the study phase and make an old vs new decision by button press. The task had the following outcomes: (1) a studied word was correctly classified as old or 'hit' (true memory recognition); (2) a studied word was incorrectly classified as new or 'miss'; (3) a non-studied word was correctly classified as new or 'correct rejection of new word'; (4) a non-studied word was incorrectly classified as old or 'false alarm'; (5) a lure was correctly classified as new or 'false memory rejection'; and (6) a lure was incorrectly classified as old or 'false recognition'.

Functional magnetic imaging protocol

Data were acquired in a 3-Tesla Siemens Magnetom Trio Scanner. Structural images of the brain were obtained by means of a T1-weighted MPRAGE sequence: 256×256 matrix; 240 1-mm sagittal slices. Functional images were obtained using an echo-planar-imaging sequence. The pulse-sequence parameters were as follows: time to repeat = 2000 ms; time to echo = 29 ms; flip angle = 80°; matrix = 128×128; slice thickness = 4 mm. Each volume comprised 36 transversal slices ($2\times2\times4$ mm voxel). A total of 412 volumes were acquired during the test phase.

Preprocessing of imaging data

fMRI data were analyzed using the SPM8 software. Raw echo-planar-imaging images were slice time and motion corrected. Echo-planar-imaging images were then co-registered to each individual's structural T1 image. T1 images were normalized to the T1 Montreal Neurologic Institute template and the obtained parameters were used to transform the echo-planar-imaging images into Montreal Neurologic Institute space. Normalized images were subjected to high-pass temporal filtering (128 s or 0.008 Hz) and to spatial smoothing using an 8 mm Gaussian filter.

Statistical analysis

A first-level analysis was performed for each individual using a design matrix that included the following predictors: 'hit', 'miss', 'correct rejection of new word', 'false alarm', 'false memory rejection', 'false recognition'. Motion correction parameters and the temporal and hemodynamic response function dispersion derivatives were introduced in the model as covariates. The contrast of interest 'false memory rejection' > 'correct rejection of new word' was calculated for each participant.

The second level analysis involved a between-groups (cannabis and controls) comparison using an independent-samples *t*-test for the 'false memory rejection' > 'correct rejection of new word' contrast. Both the controls > cannabis and cannabis > controls contrasts were calculated. We considered clusters to be significantly different between groups for *P*-values <0.001 uncorrected and a spatial extension of 10 contiguous voxels.

To assess for correlations between activation values and drug-use variables, mean fMRI parameter values for the different statistically significant clusters (region of interest) were calculated for each individual. The voxels included in the calculations for each cluster were those showing *P*-values <0.001 uncorrected.

RESULTS

Behavior

The analysis of behavioral data obtained in the study phase did not detect differences between groups regarding their degree of attention. Thus, the number of correctly identified catch trials, expressed as mean ± s.d., was 4.00 ± 0.63 for the controls and 4.18 ± 0.75 for the cannabis users $t(30)=-0.76$, $P>0.1$.

The analysis of behavioral data in the test phase showed no differences between groups in the number of correctly recognized studied words (true memory recognition; mean ± s.d.: cannabis users, 64 ± 6; controls, 65 ± 6; $t(30) = 0.4$, $P > 0.1$) or in the number of correctly rejected new words (correct rejection of new words: cannabis users, 37 ± 3; controls, 39 ± 0.7; $t(30) = 1.9$, $P = 0.076$). No differences were found either in the time (in milliseconds) taken to correctly recognize studied words (cannabis users, 1185 ± 199; controls, 1089 ± 195; $t(30) = -1.36$, $P > 0.1$), or to correctly reject new words (cannabis users, 1200 ± 345; controls, 1043 ± 196; $t(30) = -1.58$, $P > 0.1$). However, as shown in Figure 1, cannabis users showed significantly more false memories. A two-way analysis of variance, with outcome (false recognition vs false memory rejection) as within-subjects factor and participant group (cannabis vs controls) as between-subject factors, showed a significant interaction ($F(1,30) = 5.60$, $P = 0.025$). Lure words were falsely recognized as studied words more often (false recognition; cannabis users, 12 ± 6; controls, 8 ± 4; $t(30) = -2.24$, $P = 0.033$), and were rejected less often (false memory rejection; cannabis users, 27 ± 6; controls, 32 ± 4; $t(30) = 2.46$, $P = 0.021$).

fMRI

Imaging data were analyzed specifically looking for differences between groups in the pattern of blood oxygenation level dependent (BOLD) response associated with the correct rejection of lures or false memory rejection as compared with the correct rejection of new words. Figure 2 shows the mentioned contrast separately for each of the two participant groups. Note the larger extension and lower *P*-values of active voxels in the control group.

Figure 3 and Table 1 show the results of the between-groups comparison. Control participants showed higher activation for the contrast false memory rejection > correct rejection of new words in parietal, prefrontal, temporal and subcortical structures. All these structures have previously been found to be involved in the correct identification of false relative to new semantic stimuli.[26] Greater behavioral efficacy in the control group was thus associated with greater brain activity for the rejection of lures than for the rejection of new unrelated words.

Correlation analysis

To look for potential associations between the pattern of brain activation and history of cannabis use, we defined regions of interest for each of the statistically significant areas identified in the between-groups comparison. The parameters (beta values) associated with false memory rejection in each region of interest were extracted only for the cannabis group, and their values were correlated with drug-use data: lifetime cannabis consumption, years of use and amount of cannabis used daily. As shown in Figure 4, a significant negative correlation ($r = -0.806$, $r^2 = 0.650$, $P < 0.001$) was found between activity in the MTL regions of interest and lifetime cannabis use (log value of the estimated number of cannabis cigarettes smoked).

DISCUSSION

Our results show that cannabis users had a higher susceptibility to memory illusions, as observed in certain neurologic and psychiatric populations,[21] and elderly individuals.[23] They further identify the functional substrate of this deficit in the hypoactivation of a series of spatially distributed brain regions participating in the network involved in semantic[30] and episodic[31] retrieval. The network identified fits nicely with previous studies that have shown that compared with new items, recognition of false stimuli leads to greater activation of the hippocampus and the parahippocampal gyrus, and also of the left parietal and left dorsolateral prefrontal cortices in healthy subjects.[26] Although activation of MTL structures in these tasks can be directly

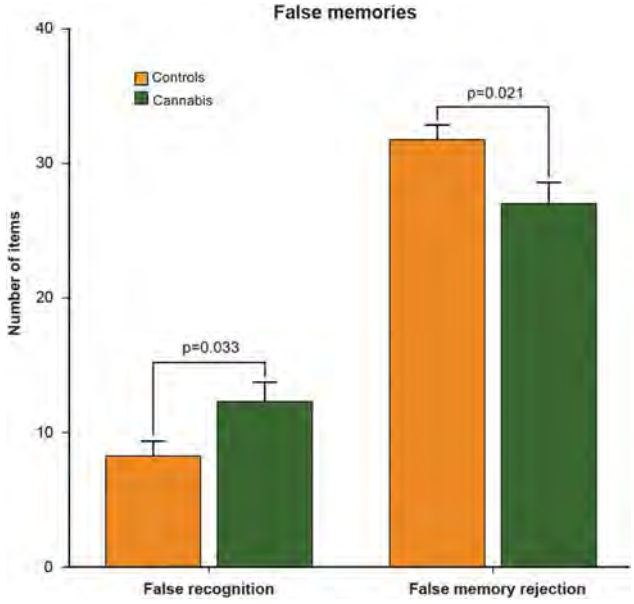


Figure 1. Behavioral data. The graphs show performance results in the memory task. Cannabis users performed significantly worse than controls, showing increased false recognition and decreased false memory rejection. Error bars denote one s.d. of mean.

associated with memory,[32] the parietal cortex can be linked to attentional processes and the dorsolateral prefrontal cortex to monitoring issues in this context.[33] It has been shown that the effective rejection of lures leads to greater activation of the dorsolateral prefrontal cortex[34] and lesions at this level lead to increased false recognition in neurological patients.[35] Thus, rather than a compromise of memory structures per se (that is, the hippocampus), our results point to a more diffuse impairment, which leads to a reduced capacity to deal with the retrieval and monitoring demands needed to differentiate between illusory and real events.

From a theoretical perspective, two main accounts have been put forward to explain the false memory phenomenon: the fuzzy trace theory and the activation-monitoring account. The fuzzy trace theory postulates that stimuli are encoded into two types of memory traces: a 'verbatim' trace containing specific details and features associated with the stimulus, and a 'gist' trace that contains more general aspects of the encoding event. False memories occur when new stimuli share certain features with past events and elicit the retrieval of the gist trace, but not the verbatim trace.[36] In contrast, the activation-monitoring account[37] postulates that cognitive control mechanisms need to be engaged to correctly identify and reject the highly activated lures. According to this view, false memories occur when monitoring mechanisms fail to identify the non-studied but semantically related lures.

Our findings can be interpreted in the light of the two accounts described above. The between-groups comparison of fMRI activation maps showed activity not only in distributed brain areas participating in semantic[30] and episodic[31] retrieval, but also in cognitive control, as suggested by the significant dorsolateral prefrontal clusters identified.[38,39] The greater activation found for the control group in the medial and lateral temporal cortices suggests access to both the semantic (lateral) and episodic (medial) features of the studied stimuli. Using the terminology of the fuzzy trace account, controls would take advantage of both the verbatim and gist traces when deciding to reject a false memory. On the contrary, the inverse correlation found between lifetime cannabis use and the BOLD response in the MTL suggests that chronic exposure to cannabis may be especially detrimental to the brain structure providing the episodic or gist features to

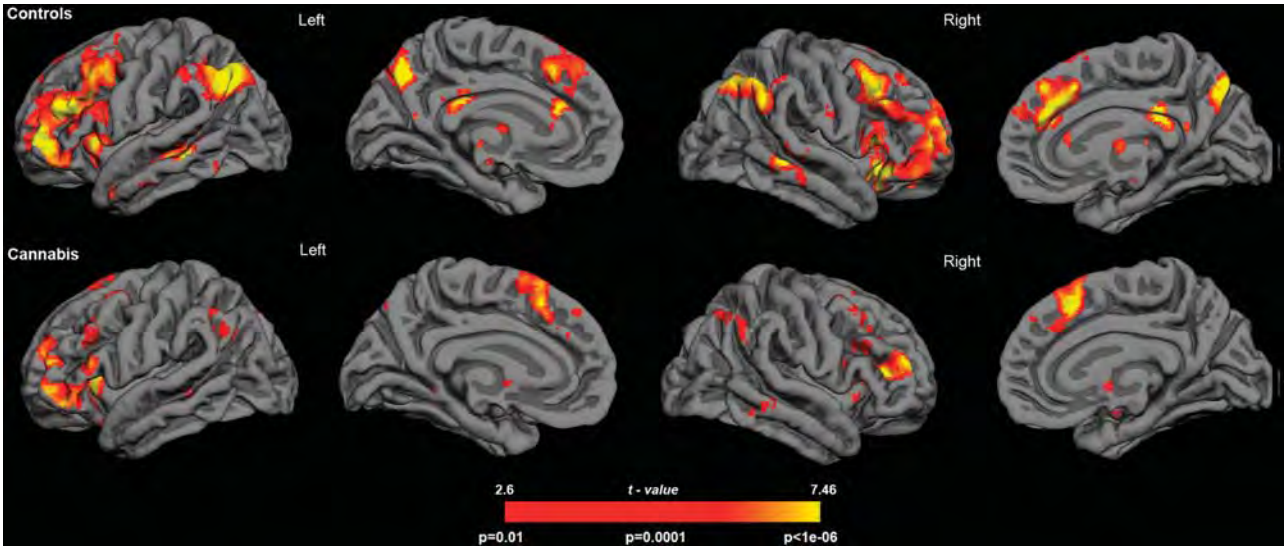


Figure 2. Rendering of fMRI results for each participant group. The statistical maps show the results of the voxel-wise comparison 'false memory rejection' > 'correct rejection of new word'. For depiction purposes results are shown at $P = 0.01$.

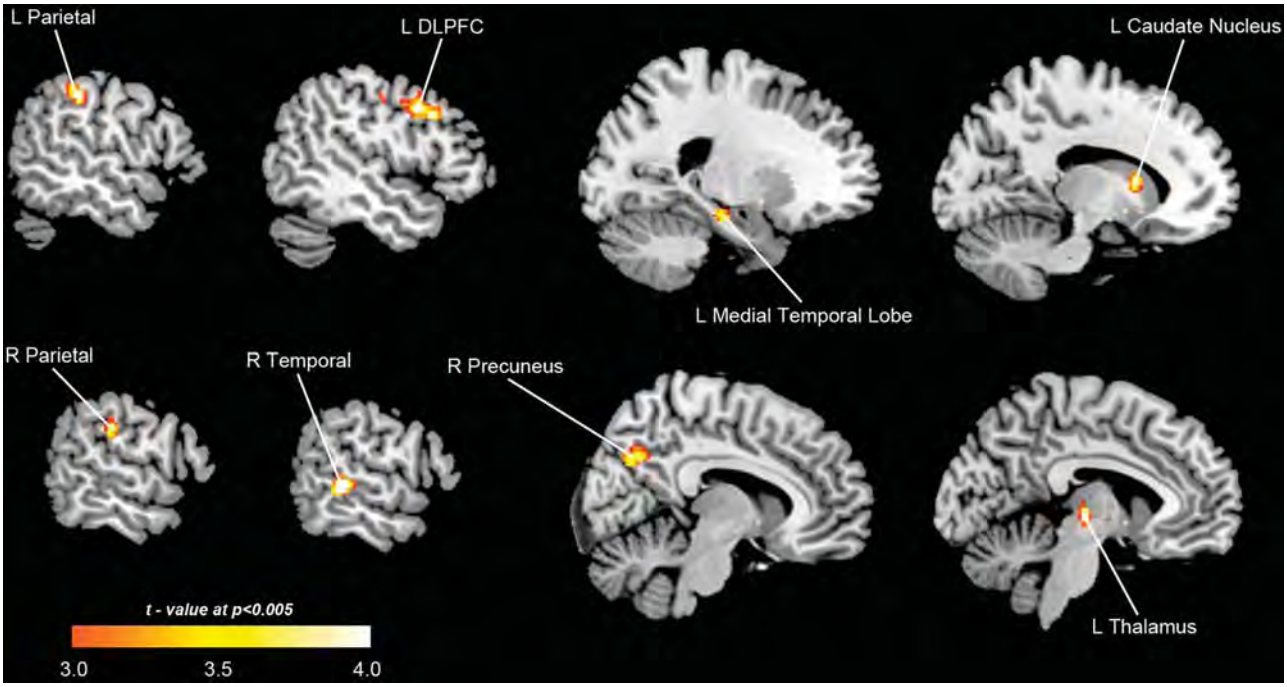


Figure 3. Group differences between controls and cannabis users. The images show the results of the voxel-wise independent-samples *t*-test controls > cannabis users for the contrast 'false memory rejection' > 'correct rejection of new word'. The brain regions depicted showed significantly higher activation in the control group as compared with the cannabis using group at $P = 0.001$ uncorrected. No significant results were obtained for the contrast cannabis users > controls. For depiction purposes results are shown at $P = 0.005$.

stored information. Cannabis users may have been left more dependent on the verbatim features of stimuli to decide whether a given word was a legitimate memory or not. Paradoxically, the greater activation of gist-related information in the control group compared with the cannabis group might have made them more vulnerable to false memories. Concurrent retrieval of item-based (verbatim) and context-based (gist) information in the control group might elicit conflict and require the engagement of cognitive control mechanisms, explaining the increased frontal activation observed in the controls. Thus, a more efficient conflict- or activation-monitoring, as signaled by increased dorsolateral prefrontal activity, may have led to the final outcome of better performance in the control group.

Further evidence of MTL and prefrontal impairment by cannabis is provided by magnetic resonance spectroscopy studies. Using this technique, researchers have found detrimental neurometabolic changes in these brain areas. For instance, Silveri and colleagues have reported decreased myo-inositol/creatine levels in the MTL and thalamus of users.[40,41] Hermann *et al.*[42] have found reduced *N*-acetyl-aspartate/total creatine ratios in the dorsolateral prefrontal of recreational users, and Cowan *et al.*[43] have found analogous decreases in Brodmann area 45 in the inferior frontal

Table 1. Areas of increased BOLD response in controls relative to cannabis users for the contrast: 'false memory rejection' > 'correct rejection of new word'

Brain region	*Lateralization*	*BA*	*MNI (x, y, z)*	*Maximum* t	n *voxels*
Temporal cortex	Right	22	60, −34, 2	5.14	64
Dorsolateral prefrontal	Left	9	−52, 16, 34	5.06	76
Red nucleus/thalamus	Left	–	−6, −18, −2	4.76	24
Parietal cortex	Left	40	-56, −32, 42	4.66	71
Parietal cortex	Right	40	60, −26, 28	4.03	15
Caudate	Left	–	−14, 12, 10	3.90	20
Medial temporal lobe	Left	35/28	−22, −22, −16	3.87	17
Precuneus	Right	7	10, −70, 32	3.82	23

Abbreviations: BA, Brodmann area; BOLD, blood oxygenation level dependent; MNI, coordinates in Montreal Neurological Institute stereotactic space; *t, t*-value (df = 30).

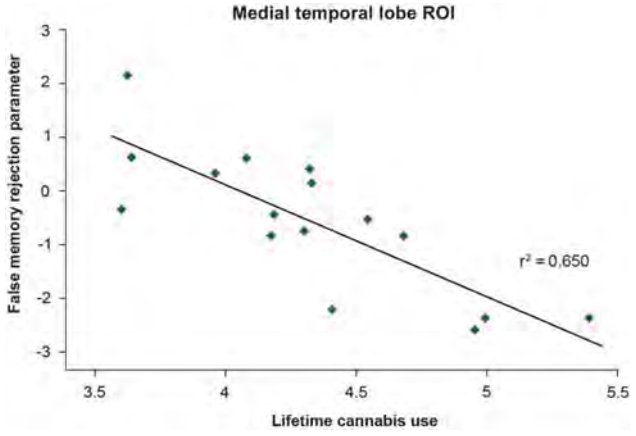


Figure 4. Correlation between MTL activity and cannabis exposure. The scatter plot shows the relationship between the individual statistical parameters in the MTL (beta values) associated with the 'false memory rejection' condition and lifetime cannabis use (log of estimated total number of cannabis cigarettes).

gyrus. Considering that analogous neurometabolic changes can be observed in older individuals[44] and that reality monitoring deficits increase with age,[45] we speculate that chronic cannabis use could aggravate the memory deficits associated with the normal ageing process.

Our findings extend previous knowledge on the impact of cannabis use on memory[12] and executive function.[8] Although there are contradictory results regarding the normalization of memory in the long term,[9,10,16] impairment has been associated with the intensity of cannabis use, with heavy users showing deficits in various memory functions.[46] Interestingly, many neuroimaging studies implementing simple memory tasks have failed to find differences in performance between heavy cannabis users and controls.[12] Our findings suggest that impairment may be more subtle and affect more complex cognitive processes, like those involved in the Deese-Roediger-McDermott paradigm.

A limitation of our study is the potential presence of residual THC levels in the brain in the absence of detectable levels in other biological matrices (in our case, urine). Whereas most studies in humans consider that cognitive testing after a 4-week period will assess the long-term effects of cannabis rather than its residual effects,[8] a longer persistence of THC in the brain has also been observed.[47] Thus, although unlikely, the presence of small amounts of THC in the body cannot be entirely ruled out.

Taken together, the present results indicate that long-term heavy cannabis users are at an increased risk of experiencing memory errors even when abstinent and drug-free. These deficits show a neural basis and suggest a subtle compromise of brain mechanisms involved in reality monitoring. Though subtle, the deficits found bear similarities with alterations observed in psychiatric and neurologic conditions and also with age-related cognitive decline. This lingering diminished ability to tell true from false may have medical, and legal implications. Future studies should address these issues and assess whether the deficits found here extend to other forms of memory distortion and reality monitoring beyond the false memory phenomenon.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

This work was supported by a grant from the 'Plan Nacional Sobre Drogas' of the Spanish Government. Marta Valle is supported by the 'Fondo de Investigación Sanitaria' through grant CP04/00121 from the Spanish Ministry of Health in collaboration with Institut de Recerca de l'Hospital de la Santa Creu i Sant Pau, Barcelona. Frederic Sampedro is supported by an FPU grant from the Spanish government. The authors wish to thank José Carlos Bouso for his help during participant recruitment and data collection, and Cesar Garrido and Núria Bargalló for technical assistance.

REFERENCES

1 United Nations Office on Drugs and Crime. World Drug Report 2014 2014; http://www.unodc.org/wdr2014/ (accessed 9 Mar2014).
2 Substance Abuse and Mental Health Services Administration. The NSDUH Report: Substance Use and Mental Health Estimates from the 2013 National Survey on Drug Use and Health: Overview of Findings 2014.
3 Wade L. South America. Legal highs make Uruguay a beacon for marijuana research. *Science* 2014; **344**: 1217.
4 Volkow ND, Baler RD, Compton WM. Weiss SRB. Adverse Health Effects of Marijuana Use. *N Engl J Med* 2014; **370**: 2219–2227.
5 Lopez-Quintero C, Pérez de los Cobos J, Hasin DS, Okuda M, Wang S, Grant BF *et al.* Probability and predictors of transition from first use to dependence on nicotine, alcohol, cannabis, and cocaine: results of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). *Drug Alcohol Depend* 2011; **115**: 120–130.
6 Patton GC, Coffey C, Carlin JB, Degenhardt L, Lynskey M, Hall W. Cannabis use and mental health in young people: cohort study. *BMJ* 2002; **325**: 1195–1198.
7 Caspi A, Moffitt TE, Cannon M, McClay J, Murray R, Harrington H *et al.* Moderation of the effect of adolescent-onset cannabis use on adult psychosis by a functional polymorphism in the catechol-O-methyltransferase gene: longitudinal evidence of a gene X environment interaction. *Biol Psychiatry* 2005; **57**: 1117–1127.
8 Crean RD, Crane NA, Mason BJ. An evidence based review of acute and long-term effects of cannabis use on executive cognitive functions. *J Addict Med* 2011; **5**: 1–8.
9 Meier MH, Caspi A, Ambler A, Harrington H, Houts R, Keefe RSE *et al.* Persistent cannabis users show neuropsychological decline from childhood to midlife. *Proc Natl Acad Sci USA* 2012; **109**: E2657–E2664.
10 Solowij N, Battisti R. The chronic effects of cannabis on memory in humans: a review. *Curr Drug Abuse Rev* 2008; **1**: 81–98.

11 Solowij N, Jones KA, Rozman ME, Davis SM, Ciarrochi J, Heaven PCL *et al.* Verbal learning and memory in adolescent cannabis users, alcohol users and non-users. *Psychopharmacology (Berl)* 2011; **216**: 131–144.
12 Schoeler T, Bhattacharyya S. The effect of cannabis use on memory function: an update. *Subst Abuse Rehabil* 2013; **4**: 11–27.
13 Ranganathan M, D'Souza DC. The acute effects of cannabinoids on memory in humans: a review. *Psychopharmacology (Berl)* 2006; **188**: 425–444.
14 Curran HV, Brignell C, Fletcher S, Middleton P, Henry J. Cognitive and subjective dose-response effects of acute oral Delta 9-tetrahydrocannabinol (THC) in infrequent cannabis users. *Psychopharmacology (Berl)* 2002; **164**: 61–70.
15 Englund A, Morrison PD, Nottage J, Hague D, Kane F, Bonaccorso S *et al.* Cannabidiol inhibits THC-elicited paranoid symptoms and hippocampal-dependent memory impairment. *J Psychopharmacol* 2013; **27**: 19–27.
16 Pope HG, Gruber AJ, Hudson JI, Huestis MA, Yurgelun-Todd D. Neuropsychological performance in long-term cannabis users. *Arch Gen Psychiatry* 2001; **58**: 909–915.
17 Pope HG, Gruber AJ, Hudson JI, Huestis MA, Yurgelun-Todd D. Cognitive measures in long-term cannabis users. *J Clin Pharmacol* 2002; **42**: 41S–47S.
18 Ashtari M, Avants B, Cyckowski L, Cervellione KL, Roofeh D, Cook P *et al.* Medial temporal structures and memory functions in adolescents with heavy cannabis use. *J Psychiatr Res* 2011; **45**: 1055–1066.
19 Cousijn J, Wiers RW, Ridderinkhof KR, van den Brink W, Veltman DJ, Goudriaan AE. Grey matter alterations associated with cannabis use: results of a VBM study in heavy cannabis users and healthy controls. *Neuroimage* 2012; **59**: 3845–3851.
20 Yücel M, Solowij N, Respondek C, Whittle S, Fornito A, Pantelis C *et al.* Regional brain abnormalities associated with long-term heavy cannabis use. *Arch Gen Psychiatry* 2008; **65**: 694–701.
21 Kopelman MD. Varieties of false memory. *Cogn Neuropsychol* 1999; **16**: 197–214.
22 Schacter DL. The seven sins of memory. Insights from psychology and cognitive neuroscience. *Am Psychol* 1999; **54**: 182–203.
23 Dennis NA, Bowman CR, Peterson KM. Age-related differences in the neural correlates mediating false recollection. *Neurobiol Aging* 2014; **35**: 395–407.
24 Fuentemilla L, Càmara E, Münte TF, Krämer UM, Cunillera T, Marco-Pallarés J *et al.* Individual differences in true and false memory retrieval are related to white matter brain microstructure. *J Neurosci* 2009; **29**: 8698–8703.
25 Roediger HL, McDermott KB. Creating false memories: remembering words not presented in lists. *J Exp Psychol Learn Mem Cogn* 1995; **24**: 803–814.
26 Cabeza R, Rao SM, Wagner AD, Mayer AR, Schacter DL. Can medial temporal lobe regions distinguish true from false? An event-related functional MRI study of veridical and illusory recognition memory. *Proc Natl Acad Sci USA* 2001; **98**: 4805–4810.
27 Nelson HE, O'Connell A. Dementia: the estimation of premorbid intelligence levels using the New Adult Reading Test. *Cortex* 1978; **14**: 234–244.
28 Del Ser T, González-Montalvo JI, Martínez-Espinosa S, Delgado-Villapalos C, Bermejo F. Estimation of premorbid intelligence in Spanish people with the Word Accentuation Test and its application to the diagnosis of dementia. *Brain Cogn* 1997; **33**: 343–356.
29 Wechsler D. *Wechsler Adult Intelligence Scale-III (WAIS-III)*. The Psychological Corporation: San Antonio, TX, USA, 1981.
30 Binder JR, Desai RH, Graves WW, Conant LL. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb Cortex* 2009; **19**: 2767–2796.
31 Sheldon S, Moscovitch M. The nature and time-course of medial temporal lobe contributions to semantic retrieval: an fMRI study on verbal fluency. *Hippocampus* 2012; **22**: 1451–1466.
32 Ritchey M, Wing EA, LaBar KS, Cabeza R. Neural similarity between encoding and retrieval is related to memory via hippocampal interactions. *Cereb Cortex* 2013; **23**: 2818–2828.
33 Schacter DL, Slotnick SD. The cognitive neuroscience of memory distortion. *Neuron* 2004; **44**: 149–160.
34 McDermott KB, Jones TC, Petersen SE, Lageman SK, Roediger HL. Retrieval success is accompanied by enhanced activation in anterior prefrontal cortex during recognition memory: an event-related fMRI study. *J Cogn Neurosci* 2000; **12**: 965–976.
35 Parkin AJ, Bindschaedler C, Harsent L, Metzler C. Pathological false alarm rates following damage to the left frontal cortex. *Brain Cogn* 1996; **32**: 14–27.
36 Brainerd CJ, Reyna VF. Fuzzy-trace theory and false memory. *Curr Dir Psychol Sci* 2002; **11**: 164–169.
37 Balota DA, Cortese MJ, Duchek JM, Adams D, Roediger HL, McDermott KB *et al.* Veridical and false memories in healthy older adults and in dementia of the Alzheimer's type. *Cogn Neuropsychol* 1999; **16**: 361–384.
38 Ridderinkhof KR, Ullsperger M, Crone EA, Nieuwenhuis S. The role of the medial frontal cortex in cognitive control. *Science* 2004; **306**: 443–447.
39 Petrides M. Lateral prefrontal cortex: architectonic and functional organization. *Philos Trans R Soc B Biol Sci* 2005; **360**: 781–795.
40 Mashhoon Y, Jensen JE, Sneider JT, Yurgelun-Todd DA, Silveri MM. Lower Left Thalamic Myo-Inositol. Levels Associated with Greater Cognitive Impulsivity in Marijuana-Dependent Young Men: Preliminary Spectroscopic Evidence at 4T. *J Addict Res Ther* 2013; doi: 10.4172/2155-6105.S4-009.
41 Silveri MM, Jensen JE, Rosso IM, Sneider JT, Yurgelun-Todd DA. Preliminary evidence for white matter metabolite differences in marijuana-dependent young men using 2D J-resolved magnetic resonance spectroscopic imaging at 4 Tesla. *Psychiatry Res* 2011; **191**: 201–211.
42 Hermann D, Sartorius A, Welzel H, Walter S, Skopp G, Ende G *et al.* Dorsolateral prefrontal cortex N-acetylaspartate/total creatine (NAA/tCr) loss in male recreational cannabis users. *Biol Psychiatry* 2007; **61**: 1281–1289.
43 Cowan RL, Joers JM, Dietrich MS. N-acetylaspartate (NAA) correlates inversely with cannabis use in a frontal language processing region of neocortex in MDMA (Ecstasy) polydrug users: a 3T magnetic resonance spectroscopy study. *Pharmacol Biochem Behav* 2009; **92**: 105–110.
44 Fukuzako H, Hashiguchi T, Sakamoto Y, Okamura H, Doi W, Takenouchi K *et al.* Metabolite changes with age measured by proton magnetic resonance spectroscopy in normal subjects. *Psychiatry Clin Neurosci* 1997; **51**: 261–263.
45 McDaniel MA, Lyle KB, Butler KM, Dornburg CC. Age-related deficits in reality monitoring of action memories. *Psychol Aging* 2008; **23**: 646–656.
46 Bolla KI, Brown K, Eldreth D, Tate K, Cadet JL. Dose-related neurocognitive effects of marijuana use. *Neurology* 2002; **59**: 1337–1343.
47 Mura P, Kintz P, Dumestre V, Raul S, Hauet T. THC can be detected in brain while absent in blood. *J Anal Toxicol* 2005; **29**: 842–843.

Supplementary Information accompanies the paper on the Molecular Psychiatry website (http://www.nature.com/mp)