Automated assessment of FDG-PET for the differential diagnosis in patients with neurodegenerative disorders

Flavio Nobili*, MD; Cristina Festari, MS; Daniele Altomare, MS; Federica Agosta, MD, PhD; Stefania Orini, MD; Koen Van Laere, MD PhD DSc; Javier Arbizu, MD; Femke Bouwman, MD; Alexander Drzezga, MD; Peter Nestor, MD; Zuzana Walker, MD; Marina Boccardi*, PhD; *for the EANM-EAN Task Force for the Prescription of FDG-PET for Dementing Neurodegenerative Disorders.*

AFFILIATIONS

Flavio Nobili: Dept. of Neuroscience (DINOGMI) University of Genoa and Polyclinic San Martino Hospital, Genoa, Italy.

Cristina Festari: LANE – Laboratory of Alzheimer's Neuroimaging & Epidemiology, IRCCS S. Giovanni di Dio, Fatebenefratelli, Brescia, Italy; Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy.

Daniele Altomare: LANE – Laboratory of Alzheimer's Neuroimaging & Epidemiology, IRCCS S. Giovanni di Dio, Fatebenefratelli, Brescia, Italy; Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy.

Federica Agosta: Neuroimaging Research Unit, Institute of Experimental Neurology, Division of Neuroscience, San Raffaele Scientific Institute, Vita-Salute San Raffaele University, Milan, Italy.

Stefania Orini: Alzheimer Operative Unit, IRCCS S. Giovanni di Dio, Fatebenefratelli, Brescia, Italy.

Koen Van Laere: Division of Nuclear Medicine, University Hospitals Leuven and Department of Imaging and Pathology, KU Leuven, Leuven, Belgium

Javier Arbizu: Department of Nuclear Medicine, Clinica Universidad de Navarra. University of Navarra, Pamplona, Spain.

Femke Bouwman: Department of Neurology & Alzheimer Center, Amsterdam Neuroscience, VU University Medical Center, Amsterdam, the Netherlands.

Alexander Drzezga: Department of Nuclear Medicine, University Hospital of Cologne, University of Cologne and German Center for Neurodegenerative Diseases (DZNE), Germany.

Peter Nestor: German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany; Queensland Brain Institute, University of Queensland and Mater Hospital, Brisbane, Australia.

Zuzana Walker: University College London, Division of Psychiatry & Essex Partnership University.

Marina Boccardi: LANVIE (Laboratoire de Neuroimagerie du Vieillissement), Dept of Psychiatry, University of Geneva, Geneva, Switzerland; LANE – Laboratory of Alzheimer's Neuroimaging & Epidemiology, IRCCS S. Giovanni di Dio, Fatebenefratelli, Brescia, Italy

*Correspondents:
Flavio Nobili
Dept. of Neuroscience (DINOGMI)
University of Genoa
And Clinical Neurology
Polyclinic IRCCS San Martino Hospital
Genoa, Italy
e-mail: flaviomariano.nobili@hsanmartino.it
Tel.: 0039.010.3537568
Fax.: 0039.010.5556893

Marina Boccardi
LANVIE (Laboratoire de Neuroimagerie du Vieillissement), Dept of Psychiatry
University of Geneva, Geneva, Switzerland
Chemin du Petit-Bel-Air, 2
1225, Chene-Bourg, Genève
e-mail: marina.boccardi@unige.ch
Tel.: 0041.(0)22.3055764
Fax.: 0041.(0)22.3054719

ABSTRACT

Purpose: To review literature until November 2015 and reach a consensus on whether automatic semi-quantification of brain FDG-PET is useful in the clinical setting for neurodegenerative disorders.

Methods: Literature search was conducted in Medline, Embase, and Google scholar. Papers were selected with a lower limit of 30 patients (no limits with autopsy confirmation). Consensus recommendations were developed through a Delphi procedure, based on the expertise of panelists, who were also informed about the availability and quality of evidence, assessed by an independent methodology team.

Results: Critical outcomes were available in nine among the 17 papers initially selected. Only three papers performed the direct comparison between visual and automated assessment and quantified the incremental value provided by the latter. Sensitivity between visual and automatic analysis is similar but automatic assessment generally improves specificity, and marginally accuracy. Also, automated assessment increases diagnostic confidence. As expected, performance of visual analysis is reported to depend on expertise of readers.

Conclusions: Tools for semi-quantitative evaluation are recommended to assist the nuclear medicine physician in reporting brain FDG-PET pattern in neurodegenerative conditions. However, heterogeneity, complexity and drawbacks of these tools should be known by users to avoid misinterpretation. Head-to-head comparisons and an effort to harmonize procedures are encouraged.

INTRODUCTION

Clinical guidelines for FDG-PET in the diagnosis of dementia are lacking, thus the European Association of Nuclear Medicine (EANM) and the European Academy of Neurology (EAN) performed a joint project to provide guidance to clinicians in using of the exam, namely for 20 clinical situations. These have been addressed based on literature evidence and expert consensus [1]. In addition to the clinical issues, a 21st question was defined, relating to the opportunity to assist the traditional visual reading with automatic (or semi-automatic) semi-quantification of FDG-PET images. In this paper, we report our assessment of the literature on this topic, and specifically of accuracy studies supporting the use of automated assessment, as provided to the panelist taking decisions in a Delphi panel, and detail the reasons for their decisions.

Evaluation of FDG-PET images is minimally performed through visual reading on the three space planes by a nuclear medicine physician, as suggested by the EANM procedure guidelines [2], the standard procedure to report FDG-PET. In some European countries the tridimensional image view is also requested. As for other morphological or functional imaging modalities, considerable training and expertise are required to achieve high-quality reporting especially in clinical conditions with subtle defects or tricky findings, as sometimes happens in the earliest stages of most neurodegenerative disorders. In these conditions, automatic tools for semi-quantification can be helpful by assisting even the expert reader in increasing confidence in diagnostic conclusion. However, even such automatic systems have their own peculiarities and limitations, which differ from one another, that should be well known and carefully considered by users. While aiding the reader on the one hand, on the other hand they can generate artifacts or wrong results that may be confounding. There are now several such fully or semi-automatic systems; some are free on the web but the majority are commercial. The Statistical Parametric Mapping (SPM) tool [3] is probably the most popular one since it is free and it is periodically updated. Although it has been designed for group comparisons on a voxel-by-voxel basis, it has been adapted and validated for use in

comparison between a single subject versus a control group [4,5], but is used without CE label. Moreover, a dementia-customized FDG-PET template has been made available for SPM [6]. However, a normal control group is not embedded within the tool and must be built to perform comparisons, and both the statistical thresholds and the reference region are to be chosen by the user and can be varied, with both advantages and disadvantages towards sensitivity and specificity [3]. Three-dimension statistical surface projection (3D-SSP, Neurostat®) [7] is another popular and free tool for voxel-based analysis, requires subscription, and has been specifically designed for single-case comparison versus a control group which is instead embedded within the system and can be further implemented according to the needs of the center. There are several CE products (such as Cortex-ID® and MIMVista®) and again, both thresholds and reference regions can be chosen. Commercial tools have been generated following academic implementation, such as the PALZ score [8] of the PMOD software and syngo.PET Neuro DataBase Comparison®, or have been implemented by scanner manufacturers and are included in the PET workstations. They generally follow similar rules and characteristics as the three above mentioned tools. A common point for all these tools is that the resulting statistical maps still need to be interpreted by the reader.

Other automatic methods relying on volumetric regions of interest (VOIs) have been developed by research groups and have been applied to patients with AD or MCI converting to AD dementia or not [9]. Examples of these tools are the hypometabolic convergence index (HCI) [10], the metaROIs which requires SPM sub-tools [11] , and similar other systems [12]. These approaches have all shown good accuracy but may be more time-consuming to be used in the clinical setting as they are poorly automatized.

Within the EANM-EAN project, we performed a literature search to assess the quality of the evidence on the added value of automated semi-quantification of FDG-PET scans in patients with neurodegenerative disorders, including all kinds of approaches. The results of this assessment allowed panelists to generate recommendations through a Delphi consensus procedure, on whether and when supporting the use of computer-aided tools. Only studies dealing with semi-quantification

where activity is normalized to a disease free region (or to the whole brain activity) were considered as they are the vast majority and because absolute quantification requiring long dynamic acquisition times and an invasive procedure can hardly be implemented in clinical practice.

METHODS

Seven panelists, four from EANM and three from EAN, were appointed to produce recommendations taking into consideration the opportunity to support the use of semi-automated assessment to assist visual reading in the clinical setting. Consensus recommendations have been produced through a Delphi procedure based on the expertise of panelists, who were also informed about the availability and quality of evidence, assessed by an independent methodology team as described in Boccardi et al [13].

Briefly, we performed literature searches using harmonized PICO (Population, Intervention, Comparison, Outcome) question keywords edited by the experts, screened the studies for eligibility, extracted the data to assess their methodological quality, and provided an evidence assessment consistent with the EFNS guidance [14] though specifically adapted to FDG-PET studies [13].

PICO question for this paper. For this review, the PICO question was whether *automated semiquantitative assessment of FDG-PET scans should be required, as adding sufficient information (in terms of increased accuracy, and versus pathology, biomarker-based diagnosis or conversion at follow-up) as compared to visual reading alone, to optimize the diagnostic work-up of patients with dementing neurodegenerative disorders.*

Eligibility criteria. Only original full papers published in English on international impacted journals were considered, excluding reviews, management guidelines, abstracts and gray literature. Any sample size was allowed if pathology was the gold standard for diagnosis. Otherwise, 30 subjects, demented patients and/or healthy controls, were requested as the minimum sample sizes.

Literature search. Electronic search strategy, developed and tested with panelists, was performed using predefined strings, grounding on the specific PICO question and including a selection of terms taken from a largely inclusive literature selection, in order to pick all variants for the same keyword. Other details of the literature search are reported in the methodological paper [13].

Data extraction and quality assessment. CF extracted data for this review. The quality of evidence was assessed consensually within the methodology team based on study design, gold/reference standard, FDG-PET image assessment (visual or semi-quantitative methods), risk of bias, index test imprecision, applicability, effect size, and effect inconsistency [13]. A final assessment of relative availability of evidence was formulated, keeping into account evidence availability among all of the 21 PICOs. This ranking was summarized as very poor/lacking, poor, fair or good.

RESULTS

Among the 17 papers identified and screened by the referent panelist (FN), 14 were sent to the methodology team for data extraction and assessment (Figure 1). Two papers [15,16] failed to show semiquantitative results and were not considered. Among the remaining 12, critical outcomes were available in 9 papers (see Table PICO 21; data extraction table available at (https://drive.google.com/file/d/0B0_JB3wzTvbpM0ZmVXZHX2R6c00/view?usp=sharing).

Only three papers performed the direct comparison between visual and automated assessment and quantified the incremental value provided by the latter (see Table PICO 21: Part 1, first row). Perani at al. [5] described a statistically significant change in the level of confidence using the visual assessment of SPM maps as compared to visual assessment of native orthogonal images (2.4 versus 2.07; p=0.003) in a variety of neurodegenerative conditions. Lehman [17] compared visual assessment of native images with visual assessment of 3D-SSP (projection and z-score) maps using GE CortexID® software in patients with MCI or AD dementia. They found the following values for visual assessment of native images and visual assessment of 3D-SSP projection and z-score maps, respectively: sensitivity 0.82 and 0.83, specificity 0.41 and 0.75 (p<0.01), and AUC 0.88 and 0.72 (p=0.017). They also reported a borderline significant increase in confidence with 3D-SSP (p=0.048). Again by comparing visual assessment of native images and of 3D-SSP images, Burdette et al. [18] showed a significant (p=0.043) AUC increase with the automatic tool (from 0.94 to 0.99) in AD patients.

All of the nine selected papers [5,17–23] reported a variety of measures including accuracy, sensitivity, specificity or others (AUC, PPV, NPV, LR+, LR-) of visual and automated FDG-PET assessment in identifying disease-specific hypometabolism. The results are extensively described in Table PICO 21 (Part 1 and Part 2). In general, studies with semi-quantitative and visual reading reported similar ranges of sensitivity (62.3%-96% and 59%-89.6%, respectively) and accuracy (70%-97.5% and 64.8%-89.2%) whereas specificity was greater with automated assessment than with visual reading (range: 84%-99% vs 50%-96%). As expected, the accuracy of visual reading was often found to depend on the skills and experience of the reader[17–20,22]

Four papers[6,21,23,24], based overall on 661 subjects and on good reference standard, calculated both the positive and negative likelihood ratios (LR+ and LR−). All of the analyses performed using semi-quantitative tools achieved LR+ higher than 5 and a LR− ≤0.2 (except PALZ), demonstrating diagnostic utility [25].

With reference to the above mentioned studies, VROI analysis and 3D-SSP were the most frequently used tools, followed by SPM and PALZ. Subjects included in the studies suffered from heterogeneous neurodegenerative disorders (e.g. AD, FTLD, MCI, and Vascular dementia). Risk of bias could be identified in imprecision, use of only reference standard, and applicability of the index test (Table PICO 21, part 1 and 2) as not all memory clinics/nuclear medicine departments may be able to implement these tools.

The level of evidence supporting the clinical utility of semi-automated analysis of FDG-PET was the highest among the PICOs of the entire project, and thus considered as "good". Agreement on the recommendation was achieved on Delphi Round I, 6 panelists supporting the additional use of semi-automated processing to assist visual reading in clinical settings based on its utility both for experienced and non-experienced readers.

DISCUSSION

We assessed the evidence on the utility of computer-aided tools for semi-quantitative assessment of FDG-PET in supporting the nuclear medicine physician to achieve the best interpretation of findings in neurodegenerative disorders.

The main message coming from literature review and expert consensus is that while sensitivity is similar to the one of visual interpretation of native images in this context of differentiation of neurodegenerative disorders, specificity could be better, especially if compared to non-expert readers [21]. They can get confirmation on the intensity and statistical weight of the visually assessed abnormalities, and can thus disambiguate doubtful areas of abnormalities. Indeed, a limited experience on the wide range of findings in normal people could lead to both over- and underestimation of abnormal scans. Another common situation is the presence of symmetric,

moderate hypometabolism in paramedian brain structures, such as the cingulate gyrus, the thalami, or the medial frontal gyrus, that might be missed during visual analysis. Normally these regions have a higher metabolic level compared to the rest of the cortical ribbon, thus early and symmetric hypometabolism can be missed because although the region is hypometabolic, it looks similar to the rest of the cortical ribbon. In fact, the human eye strongly relies on concomitant evaluation of the same structure in the two hemispheres during reporting (i.e., asymmetry evaluation) and thus such findings may be difficult to unveil (Figure 2). Indeed, it has been acknowledged that even the expert readers may benefit from semi-quantitative assessment, mainly by increasing specificity and diagnostic confidence [21].

Other advantages of semi-quantitative assessment are a better standardization of repeated examinations (e.g. for therapy control and follow-up studies) and the option to generate structured reporting (e.g. by reporting regional z-scores) and easier illustration of the findings to the referring physicians, which is especially true for surface rendering approach systems.

A logical consequence of the above reasoning is that automatic semi-quantitative assessment should assist the nuclear medicine physician in interpretation of findings to increase specificity and/or sensitivity and diagnostic confidence. While providing a valuable help, it cannot replace anyway his/her visual analysis, a misunderstanding that may create the false impression that achieving expertise is unnecessary given the availability of an automatic tool. This misunderstanding should be avoided for a variety of reasons. First, the output of the majority of automatic tools is still an image, that is different from native images because it shows voxels of statistically significant lower uptake, normalized on a reference region as compared to a control group, instead of FDG uptake in a color scale. Moreover, stereotactic normalization (warping) is usually performed before statistical comparison changing individual anatomy and thus artificially modifying size and precise location of lesions. In turn, these maps still need to be interpreted, similarly as native images. Thus, the accurate knowledge of hypometabolic patterns of pathological conditions remains the pre-requisite to interpret these statistical maps exactly as for native images.

If a reader does not know, let's say, the typical findings in semantic dementia, the statistical map *per se* cannot teach him/her. Second, those tools producing a score or an index with established cut-offs will never inform about the underlying pathology, but require that the differential diagnosis is restricted *a priori* to no more than two categories, e.g., AD or normal control. This is the case, for instance, for the PALZ score or the HCI which can be used only if the differential diagnosis is between an AD patient and a normal control. Third, with the most popular free software, such as SPM and 3D-SSP, both the reference normalization region and the statistical threshold can be changed which can produce different results for the same scan. For instance, 3D-SSP can use the whole brain, pons, thalami, and cerebellum; SPM uses as a default the whole brain but the normalization region can be changed or even customized to that peculiar comparison, following the Yakushev et al [26] suggestions. These steps must be known and managed with care and require at least a basic knowledge of the software and of statistical rules which is not obvious. Fourth, the standard application of some of these software tools, such as SPM and PALZ, does not correct for atrophy, thus generating maps of significant hypometabolism indifferently for true hypometabolism and for atrophy. The 3D-SSP routine (Neurostat) (https://neurostat.neuro.utah.edu/) has an inherent atrophy correction mechanism implemented as the tool searches for the maximum peak within a predefined distance orthogonally oriented across the cortical ribbon. However, the weight of atrophy may be assessed visually through coregistration of PET with the CT contextually acquired with PET for attenuation correction. Fifth, with SPM there is no standard control group that must be built-up by the user in each center, preferably using the same scanner and the same patient preparation (e.g., eyes open or closed) and reconstruction protocols as the case under analysis. 3D-SSP has a control group that can be modified by users while PALZ, CortexID® and MIMVista® have a fixed control group. Thus, the results of comparisons between a single patient and a control group change depending on the kind and quality of the control group used, besides the reference region chosen for count normalization and the statistical threshold that is chosen determining sensitivity and specificity (see above).

Finally, the results of application of these tools may vary from one another and differences may be relevant, as shown in head-to-head comparisons [9,12]. Figure 3 shows an example of good concordance between three automatic tools and visual reading, whereas Figure 4 highlights substantially different findings depending on the tool used. Thus, these tools are certainly valuable and precious to assist visual reading, but one must be aware of their heterogeneity and intrinsic way of functioning, characterized by different patterns of strengths and limitations. Care should be paid by expert nuclear physicians to train younger personnel in this perspective.

The real world at present tells us that not all nuclear medicine centers use an automatic tool to assist in reporting. Those using them have a large choice among commercial or free methods, often non-CE licensed, and thus there is an almost complete lack of harmonization among centers. This heterogeneity adds to the intrinsic heterogeneity deriving from differences among human readers, despite the teaching efforts of the EANM as well as of the National nuclear medicine societies. Finally, scanner manufacturers add in turn their part of inhomogeneity by implementing different tools in the working stations.

As a future perspective, the scientific societies in the field could promote studies that compare the performance of different tools and guide centers to implement the most accurate, cheap, and easy-to-use, hopefully supported in this action by manufacturers. These performances should always be compared with the reading of a group of experts.

Acknowledgements

Compliance with Ethical Standards.

Conflict of Interest:

Flavio Nobili: received personal fees and non-financial support from GE Healthcare, non-financial support from Eli-Lilly and grants from Chiesi Farmaceutici

Cristina Festari: declares that she has no conflict of interest

Daniele Altomare: was the recipient of the grant allocated by the European Academy of Neurology (EAN) for data extraction and evidence assessment for the present project.

Federica Agosta: is Section Editor of NeuroImage: Clinical; has received speaker fees from Biogen Idec, Novartis, and Excellence in Medical Education; and receives or has received research supports from the Italian Ministry of Health, AriSLA (Fondazione Italiana di Ricerca per la SLA), and the European Research Council. She received personal fees from Elsevier INC.

Stefania Orini: declares that she has no conflict of interest

Koen Van Laere: declares that she has no conflict of interest

Javier Arbizu: received grants from Eli-Lilly & Co, Piramal and GE Healthcare

Femke Bouwman: declares that she has no conflict of interest

Peter Nestor: declares that he has no conflict of interest

Alexander Drzezga: received grants and non-financial support from Eli-Lilly & Co, Siemens and GE Healthcare; he also received non-financial support from Piramal

Ethical approval: This article does not contain any new studies with human participants or animals performed by any of the authors. The human studies discussed herein came exclusively from previously published research articles.

Informed consent: not applicable, this is a review article. Informed consent statement is declared in each of the revised paper.

References

1. Nobili F, Arbizu J, Bouwman F, Drzezga A, Filippi M, Nestor P, et al. EAN-EANM recommendations for the use of brain 18F-Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) in neurodegenerative cognitive impairment and dementia: Delphi consensus. Eur J Neurol. (submitted)

2. Varrone A, Asenbaum S, Vander Borght T, Booij J, Nobili F, Någren K, et al. EANM procedure guidelines for PET brain imaging using [18F]FDG, version 2. Eur J Nucl Med Mol Imaging. 2009;36:2103–10.

3. Friston K, Holmes A, Worsley K, Poline J, Frith C, Frackowiak R. Statistical parametric maps in functional imaging: A general linear approach. Hum Brain Mapp. 1994;2:189–210.

4. Signorini M, Paulesu E, Friston K, Perani D, Colleluori A, Lucignani G, et al. Rapid Assessment of Regional Cerebral Metabolic Abnormalities in Single Subjects with Quantitative and Nonquantitative [18F]FDG PET: A Clinical Validation of Statistical Parametric Mapping. Neuroimage.1999;9:63–80.

5. Perani D, Della Rosa PA, Cerami C, Gallivanone F, Fallanca F, Vanoli EG, et al. Validation of an optimized SPM procedure for FDG-PET in dementia diagnosis in a clinical setting. NeuroImage Clin. 2014;6:445–54.

6. Della Rosa PA, Cerami C, Gallivanone F, Prestia A, Caroli A, Castiglioni I, et al. A Standardized [18F]-FDG-PET Template for Spatial Normalization in Statistical Parametric Mapping of Dementia. Neuroinformatics. 2014;12:575–93.

7. Minoshima S, Frey KA, Koeppe RA, Foster NL, Kuhl DE. A diagnostic approach in Alzheimer's disease using three-dimensional stereotactic surface projections of fluorine-18-FDG PET J Nucl Med. 1995;36:1238-48

8. Herholz K, Salmon E, Perani D, Baron JC, Holthoff V, Frölich L, et al. Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG PET. Neuroimage. 2002;17:302–16.

9. Caroli A, Prestia A, Chen K, Ayutyanont N, Landau SM, Madison CM, et al. Summary Metrics to Assess Alzheimer Disease-Related Hypometabolic Pattern with 18F-FDG PET: Head-to-Head Comparison. J Nucl Med. 2012;53:592–600.

10. Chen K, Ayutyanont N, Langbaum JBS, Fleisher AS, Reschke C, Lee W, et al. Characterizing Alzheimer's disease using a hypometabolic convergence index. Neuroimage. 2011;56:52–60.

11. Landau SM, Harvey D, Madison CM, Koeppe RA, Reiman EM, Foster NL, et al. Associations between cognitive, functional, and FDG-PET measures of decline in AD and MCI. Neurobiol Aging. 2011;32:1207–18.

12. Pagani M, De Carli F, Morbelli S, Öberg J, Chincarini A, Frisoni GB, et al. Volume of interest-based [18F]fluorodeoxyglucose PET discriminates MCI converting to Alzheimer's disease from healthy controls. A European Alzheimer's Disease Consortium (EADC) study. NeuroImage Clin. 2015;7:34–42.

13. Boccardi M, Festari C, Altomare D, Gandolfo F, Orini S, Nobili F, et al. Assessing accuracy diagnostic FDG-PET studies to define clinical use for dementia diagnosis. EJNMMI.

14. Leone MA, Brainin M, Boon P, Pugliatti M, Keindl M, Bassetti CL. Guidance for the preparation of neurological management guidelines by EFNS scientific task forces - revised recommendations 2012. Eur J Neurol. 2013;20:410–9.

15. Patterson JC, Lilien DL, Takalkar A, Kelley RE, Minagar A. Potential value of quantitative analysis of cerebral PET in early cognitive decline. Am J Alzheimers Dis Other Demen. 2009;23:586–92.

16. Von Borczyskowski D, Wilke F, Martin B, Brenner W, Clausen M, Mester J, et al. Evaluation of a new expert system for fully automated detection of the Alzheimer's dementia pattern in FDG PET. J Nucl Med. 2006;27:739–43.

17. Lehman VT, Carter RE, Claassen DO, Murphy RC, Lowe V, Petersen RC, et al. Visual assessment versus quantitative three-dimensional stereotactic surface projection fluorodeoxyglucose positron emission tomography for detection of mild cognitive impairment and Alzheimer disease.

Clin Nucl Med. 2012;37:721–6.

18. Burdette JH, Minoshima S, Vander Borght T, Tran DD, Kuhl DE. Alzheimer disease: improved visual interpretation of PET images by using three-dimensional stereotaxic surface projections. Radiology. 1996;198:837–43.

19. Ishii K, Kono AK, Sasaki H, Miyamoto N, Fukuda T, Sakamoto S, et al. Fully automatic diagnostic system for early- and late-onset mild Alzheimer's disease using FDG PET and 3D-SSP. Eur J Nucl Med Mol Imaging. 2006;33:575–83.

20. Kono AK, Ishii K, Sofue K, Miyamoto N, Sakamoto S, Mori E. Fully automatic differential diagnosis system for dementia with Lewy bodies and Alzheimer's disease using FDG-PET and 3D-SSP. Eur J Nucl Med Mol Imaging. 2007;34:1490–7.

21. Morbelli S, Brugnolo A, Bossert I, Buschiazzo A, Frisoni GB, Galluzzi S, et al. Visual Versus semi-quantitative analysis of 18F-FDG-PET in Amnestic MCI: An European Alzheimer's Disease Consortium (EADC) project. J Alzheimer's Dis. 2015;44:815–26.

22. Ng S, Villemagne VL, Berlangieri S, Lee S-T, Cherk M, Gong SJ, et al. Visual Assessment Versus Quantitative Assessment of 11C-PIB PET and 18F-FDG PET for Detection of Alzheimer's Disease. J Nucl Med. 2007;48:547–52.

23. Rabinovici GD, Rosen HJ, Alkalay A, Kornak J, Furst AJ, Agarwal N, et al. Amyloid vs FDG-PET in the differential diagnosis of AD and FTLD. Neurology. 2011;77:2034–42.

24. Foster NL, Heidebrink JL, Clark CM, Jagust WJ, Arnold SE, Barbas NR, et al. FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease. Brain. 2007;130:2616–35.

25. Frisoni GB, Bocchetta M, Chételat G, Rabinovici GD, De Leon MJ, Kaye J, et al. Imaging markers for Alzheimer disease: Which vs how. Neurology. 2013. p. 487–500.

26. Yakushev I, Hammers A, Fellgiebel A, Schmidtmann I, Scheurich A, Buchholz HG, et al. SPM-based count normalization provides excellent discrimination of mild Alzheimer's disease and amnestic mild cognitive impairment from healthy aging. Neuroimage. 2009;44:43–50.

27. Moher D1, Liberati A, Tetzlaff J ADPG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 2009;62:1006–12.

FIGURE LEGENDS

Figure 1. PRISMA flowchart of selected papers for PICO 21 regarding requirement of semi-automated assessment (adapted from Moher et al. [27]).

FIGURE 2. FDG-PET native images (left) of a 78 years old woman with MCI due to AD. MMSE score = 23/30; CDR= 0.5. The bilateral, significant hypometabolism of posterior cingulate cortex is easily appreciated with an automatic tool in comparison with a normal control group, corrected for age, uncorrected for atrophy (right; PALZ, clusters of significant hypometabolism in red; [8] while it can be difficult to be detected on native images.

FIGURE 3. FDG-PET native images (a, e) of a 53 years old man with MCI due to AD. MMSE score = 28/30; CDR= 0.5. On visual analysis a moderate hypometabolism of right lateral temporal and parietal cortex was found together with bilateral hypometabolism of posterior cingulate cortex. These findings were confirmed with automatic analysis by PALZ (b, f), 3D-SSP (c: lateral projections; g: medial projections; clusters of significant hypometabolism in green-yellow; [7] which also showed clusters in anterior cingulate cortex and medial temporal lobe, not found with either PALZ or visual analysis, and SPM[4] using a height threshold of p<0.01 (h) while with the more restrictive threshold of p<0.001 only the posterior cingulate hypometabolism was highlighted (d). Other details as in Figure 2.

FIGURE 4. FDG-PET native images (a, e) of a 56 years old healthy woman with doubtful MCI and a positive family history for AD. MMSE score = 30/30; CDR= 0. On visual analysis a mild-to-moderate hypometabolism of left lateral temporal and parietal cortex was observed. These findings were confirmed with automatic analysis by PALZ (b, f), also showing significant clusters in bilateral orbitofrontal cortex while 3D-SSP (c: lateral projections; g: medial projections) essentially

failed to show left temporo-parietal hypometabolism and disclosed instead some medial frontal and temporal hypometabolic clusters. SPM did not show any significant hypometabolic cluster with a height threshold of p<0.001 or p<0.01 (d) and only a very permissive threshold of p<0.05 highlighted clusters of significant hypometabolism in left lateral frontal, temporal, and parietal cortex (h). Other details as in Figure 2-3.

Table 1. PICO 21 (PART 1). Table reports the quality of evidence for each critical outcome.

PICO 21: Incremental value of automated assessment of FDG-PET compared to visual reading											
Critical outcomes	N. of papers	Sample size	Gold/ reference standard	Risk of bias	Index test imprecision	Applicability	FDG-PET assessment	Effect range (CI)	Effect assessment	Effect inconsistency	Outcome quality
Incremental value indices	3	156 Patients 157 HC	2 Diagnosis at follow-up 1 Clinical diagnosis	Serious	Serious	Not serious	Visual + Semi-quantitative (1 SPM-Maps, 2 3D-SSP)	**Study 1 (Visual vs SPM-Maps).** - Level of confidence 2.07 vs 2.4, p=0.003. **Study 2 (Visual vs 3D-SSP).** - Sensitivity: 83% (CI: 66-94%) vs 82% (CI: 62-92%), p=1.0. - Specificity: 41% (CI: 20-61%) vs 75% (CI: 52-90%), p<0.01. - AUC: 72% (CI: 55-83%) vs 88% (CI: 76-95%), p=0.017. - Mean increase in confidence rating = 0.7 (CI: 0.01-1.3), p=0.048. **Study 3 (Visual vs 3D-SSP).** - AUC: 94% (SD 0.03) vs 0.99 (SD 0.01), p = 0.043.	MODERATE	NA	LOW
Sensitivity	6	479 Patients 126 HC	1 Pathology 1 Biomarker-based diagnosis 2 Diagnosis at follow-up 2 Clinical diagnosis	Not serious	Serious	Not serious	Visual	59% (range: 43-71%)– 89.6% (CI 80-95%)	MODERATE	Serious	MODERATE
							Semi-quantitative (2 ROI, 2 3D-SSP, 1 SPM, 1 PALZ)	62.3% (CI 50-73%) – 96% (CI NA)	MODERATE	Serious	MODERATE
Specificity	6	479 Patients 126 HC	1 Pathology 1 Biomarker-based diagnosis 2 Diagnosis at follow-up 2 Clinical diagnosis	Not serious	Serious	Not serious	Visual	50% (CI NA) – 96% (range: 92-100%)	HIGH	Very serious	LOW
							Semi-quantitative (2 ROI, 2 3D-SSP, 1 SPM, 1 PALZ)	84% (CI NA) – 99% (SD 0.02)	HIGH	Not serious	HIGH
Accuracy	7	459 Patients 237 HC	1 Pathology 1 Biomarker-based diagnosis 2 Diagnosis at follow-up 3 Clinical diagnosis	Serious	Not serious	Not serious	Visual	64.8% (CI: 51-77%) – 89.2% (CI: 84–93%)	MODERATE	Serious	MODERATE
							Semi-quantitative (3 ROI, 3 3D-SSP, 1 PALZ)	70% (CI: 53-84%) – 97.5% (CI: 91-100%)	HIGH	Serious	MODERATE

Table 1. PICO 21 (PART 2).

Critical outcomes	N. of papers	Sample size	Gold/reference standard	Risk of bias	Index test imprecision	Applicability	FDG-PET assessment	Effect range (CI)	Effect assessment	Effect inconsistency	Outcome quality
								PICO 21: Incremental value of automated assessment of FDG-PET compared to visual reading			
AUC	3	155 Patients 142 HC	1 Diagnosis at follow-up 2 Clinical diagnosis	Serious	Serious	Not serious	Visual	50% (CI NA) ; 87.8 (CI NA)	MODERATE	Serious	LOW
							Semi-quantitative (1 ROI, 1 SPM, 1 3D-SSP)	67 (CI NA);96.7 (CI NA)	HIGH	Serious	LOW
PPV	3	294 Patients 167 HC	1 Pathology 1 Biomarker-based diagnosis 1 Diagnosis at follow-up	Not serious	Serious	Not serious	Visual	68% (range: 50-88%) – 87.5% (CI NA)	HIGH	Not serious	MODERATE
							Semi-quantitative (1 ROI, 1 3D-SSP, 1 PALZ)	84.2% (CI: 72–92%) – 98% (CI: 88–100%)	HIGH	Not serious	MODERATE
NPV	3	294 Patients 167 HC	1 Pathology 1 Biomarker-based diagnosis 1 Diagnosis at follow-up	Not serious	Serious	Not serious	Visual	72% (CI NA) – 92.4% (CI: 85-96%)	HIGH	Serious	MODERATE
							Semi-quantitative (1 ROI, 1 3D-SSP, 1 PALZ)	71% (CI: 58-93%) – 89% (range: 85-92%)	MODERATE	Not serious	MODERATE
LR+	4	382 Patients 279 HC	1 Pathology 1 Biomarker-based diagnosis 2 Diagnosis at follow-up	Not serious	Serious	Serious	Visual	1.55 (CI NA) – 14.8 (CI: 10.7-∞)	MODERATE	Serious	LOW
							Semi-quantitative (1 ROI, 1 SPM, 1 3D-SSP, 1 PALZ)	6.08 (CI NA) – 36.5 (CI: 21.3-∞)	HIGH	Very serious	LOW
LR-	4	382 Patients 279 HC	1 Pathology 1 Biomarker-based diagnosis 2 Diagnosis at follow-up	Not serious	Serious	Serious	Visual	0.12 (CI: 0.06-0.23) – 0.45 (CI NA)	MODERATE	Serious	LOW
							Semi-quantitative (1 ROI, 1 SPM, 1 3D-SSP, 1 PALZ)	0.03 (CI: 0.0-0.5) – 0.41 (CI: 0.31-0.55)	MODERATE	Very serious	LOW
								RELATIVE AVAILABILITY OF EVIDENCE: STRONG			

Risk of bias: assessment of the study design and other methodological features (e.g., patient selection, clinical diagnostic criteria used). Index test methods: assessment of index test methodology (e.g., technical details, image analysis methods and statistical analysis). Applicability: representativeness of the studied population and index test reproducibility in clinical practice (semi-quantitative methods correspond to 'serious' indirectness, visual + semi-quantitative methods correspond to 'not serious' indirectness, due to partial implementation of quantitation in clinical practice). Effect: lowest and highest values for each critical outcome; when more values were obtained for the same outcome, the highest was reported. Effect assessment: 51-70% low, 71-80% moderate, 81-100% high. Effect inconsistency: 'Not serious' if lowest and highest values difference was 0-20, 'serious' 21-40, 'very serious' >40. Outcome quality: summary of evidence as from all columns.

Automated assessment of FDG-PET for the differential diagnosis in patients with neurodegenerative disorders

Flavio Nobili*, MD; Cristina Festari, MS; Daniele Altomare, MS; Federica Agosta, MD, PhD; Stefania Orini, MD; Koen Van Laere, MD PhD DSc; Javier Arbizu, MD; Femke Bouwman, MD; Alexander Drzezga, MD; Peter Nestor, MD; Zuzana Walker, MD; Marina Boccardi*, PhD; *for the EANM-EAN Task Force for the Prescription of FDG-PET for Dementing Neurodegenerative Disorders.*

AFFILIATIONS

Flavio Nobili: Dept. of Neuroscience (DINOGMI) University of Genoa and Polyclinic San Martino Hospital, Genoa, Italy.

Cristina Festari: LANE – Laboratory of Alzheimer's Neuroimaging & Epidemiology, IRCCS S. Giovanni di Dio, Fatebenefratelli, Brescia, Italy; Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy.

Daniele Altomare: LANE – Laboratory of Alzheimer's Neuroimaging & Epidemiology, IRCCS S. Giovanni di Dio, Fatebenefratelli, Brescia, Italy; Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy.

Federica Agosta: Neuroimaging Research Unit, Institute of Experimental Neurology, Division of Neuroscience, San Raffaele Scientific Institute, Vita-Salute San Raffaele University, Milan, Italy.

Stefania Orini: Alzheimer Operative Unit, IRCCS S. Giovanni di Dio, Fatebenefratelli, Brescia, Italy.

Koen Van Laere: Division of Nuclear Medicine, University Hospitals Leuven and Department of Imaging and Pathology, KU Leuven, Leuven, Belgium

Javier Arbizu: Department of Nuclear Medicine, Clinica Universidad de Navarra. University of Navarra, Pamplona, Spain.

Femke Bouwman: Department of Neurology & Alzheimer Center, Amsterdam Neuroscience, VU University Medical Center, Amsterdam, the Netherlands.

Alexander Drzezga: Department of Nuclear Medicine, University Hospital of Cologne, University of Cologne and German Center for Neurodegenerative Diseases (DZNE), Germany.

Peter Nestor: German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany; Queensland Brain Institute, University of Queensland and Mater Hospital, Brisbane, Australia.

Zuzana Walker: University College London, Division of Psychiatry & Essex Partnership University.

Marina Boccardi: LANVIE (Laboratoire de Neuroimagerie du Vieillissement), Dept of Psychiatry, University of Geneva, Geneva, Switzerland; LANE – Laboratory of Alzheimer's Neuroimaging & Epidemiology, IRCCS S. Giovanni di Dio, Fatebenefratelli, Brescia, Italy

*Correspondents:

Flavio Nobili
Dept. of Neuroscience (DINOGMI)
University of Genoa
And Clinical Neurology
Polyclinic IRCCS San Martino Hospital
Genoa, Italy
e-mail: flaviomariano.nobili@hsanmartino.it
Tel.: 0039.010.3537568
Fax.: 0039.010.5556893

Marina Boccardi
LANVIE (Laboratoire de Neuroimagerie du Vieillissement), Dept of Psychiatry
University of Geneva, Geneva, Switzerland
Chemin du Petit-Bel-Air, 2
1225, Chene-Bourg, Genève
e-mail: marina.boccardi@unige.ch
Tel.: 0041.(0)22.3055764
Fax.: 0041.(0)22.3054719

ABSTRACT

Purpose: To review literature until November 2015 and reach a consensus on whether automatic semi-quantification of brain FDG-PET is useful in the clinical setting for neurodegenerative disorders.

Methods: Literature search was conducted in Medline, Embase, and Google scholar. Papers were selected with a lower limit of 30 patients (no limits with autopsy confirmation). Consensus recommendations were developed through a Delphi procedure, based on the expertise of panelists, who were also informed about the availability and quality of evidence, assessed by an independent methodology team.

Results: Critical outcomes were available in nine among the 17 papers initially selected. Only three papers performed the direct comparison between visual and automated assessment and quantified the incremental value provided by the latter. Sensitivity between visual and automatic analysis is similar but automatic assessment generally improves specificity, and marginally accuracy. Also, automated assessment increases diagnostic confidence. As expected, performance of visual analysis is reported to depend on expertise of readers.

Conclusions: Tools for semi-quantitative evaluation are recommended to assist the nuclear medicine physician in reporting brain FDG-PET pattern in neurodegenerative conditions. However, heterogeneity, complexity and drawbacks of these tools should be known by users to avoid misinterpretation. Head-to-head comparisons and an effort to harmonize procedures are encouraged.

Keywords: brain FDG-PET, visual reading, semi-quantitative assessment, computer-aided, dementia, neurodegenerative diseases.

INTRODUCTION

Clinical guidelines for FDG-PET in the diagnosis of dementia are lacking, thus the European Association of Nuclear Medicine (EANM) and the European Academy of Neurology (EAN) performed a joint project to provide guidance to clinicians in using of the exam, namely for 20 clinical situations. These have been addressed based on literature evidence and expert consensus [1]. In addition to the clinical issues, a 21st question was defined, relating to the opportunity to assist the traditional visual reading with automatic (or semi-automatic) semi-quantification of FDG-PET images. In this paper, we report our assessment of the literature on this topic, and specifically of accuracy studies supporting the use of automated assessment, as provided to the panelist taking decisions in a Delphi panel, and detail the reasons for their decisions.

Evaluation of FDG-PET images is minimally performed through visual reading on the three space planes by a nuclear medicine physician, as suggested by the EANM procedure guidelines [2], the standard procedure to report FDG-PET. In some European countries the tridimensional image view is also requested. As for other morphological or functional imaging modalities, considerable training and expertise are required to achieve high-quality reporting especially in clinical conditions with subtle defects or tricky findings, as sometimes happens in the earliest stages of most neurodegenerative disorders. In these conditions, automatic tools for semi-quantification can be helpful by assisting even the expert reader in increasing confidence in diagnostic conclusion. However, even such automatic systems have their own peculiarities and limitations, which differ from one another, that should be well known and carefully considered by users. While aiding the reader on the one hand, on the other hand they can generate artifacts or wrong results that may be confounding. There are now several such fully or semi-automatic systems; some are free on the web but the majority are commercial. The Statistical Parametric Mapping (SPM) tool [3] is probably the most popular one since it is free and it is periodically updated. Although it has been designed for group comparisons on a voxel-by-voxel basis, it has been adapted and validated for use in

comparison between a single subject versus a control group [4,5], but is used without CE label. Moreover, a dementia-customized FDG-PET template has been made available for SPM [6]. However, a normal control group is not embedded within the tool and must be built to perform comparisons, and both the statistical thresholds and the reference region are to be chosen by the user and can be varied, with both advantages and disadvantages towards sensitivity and specificity [3]. Three-dimension statistical surface projection (3D-SSP, Neurostat®) [7] is another popular and free tool for voxel-based analysis, requires subscription, and has been specifically designed for single-case comparison versus a control group which is instead embedded within the system and can be further implemented according to the needs of the center. There are several CE products (such as Cortex-ID® and MIMVista®) and again, both thresholds and reference regions can be chosen. Commercial tools have been generated following academic implementation, such as the PALZ score [8] of the PMOD software and syngo.PET Neuro DataBase Comparison®, or have been implemented by scanner manufacturers and are included in the PET workstations. They generally follow similar rules and characteristics as the three above mentioned tools. A common point for all these tools is that the resulting statistical maps still need to be interpreted by the reader.

Other automatic methods relying on volumetric regions of interest (VOIs) have been developed by research groups and have been applied to patients with AD or MCI converting to AD dementia or not [9]. Examples of these tools are the hypometabolic convergence index (HCI) [10], the metaROIs which requires SPM sub-tools [11] , and similar other systems [12]. These approaches have all shown good accuracy but may be more time-consuming to be used in the clinical setting as they are poorly automatized.

Within the EANM-EAN project, we performed a literature search to assess the quality of the evidence on the added value of automated semi-quantification of FDG-PET scans in patients with neurodegenerative disorders, including all kinds of approaches. The results of this assessment allowed panelists to generate recommendations through a Delphi consensus procedure, on whether and when supporting the use of computer-aided tools. Only studies dealing with semi-quantification

where activity is normalized to a disease free region (or to the whole brain activity) were considered as they are the vast majority and because absolute quantification requiring long dynamic acquisition times and an invasive procedure can hardly be implemented in clinical practice.

METHODS

Seven panelists, four from EANM and three from EAN, were appointed to produce recommendations taking into consideration the opportunity to support the use of semi-automated assessment to assist visual reading in the clinical setting. Consensus recommendations have been produced through a Delphi procedure based on the expertise of panelists, who were also informed about the availability and quality of evidence, assessed by an independent methodology team as described in Boccardi et al [13].

Briefly, we performed literature searches using harmonized PICO (Population, Intervention, Comparison, Outcome) question keywords edited by the experts, screened the studies for eligibility, extracted the data to assess their methodological quality, and provided an evidence assessment consistent with the EFNS guidance [14] though specifically adapted to FDG-PET studies [13].

PICO question for this paper. For this review, the PICO question was whether *automated semiquantitative assessment of FDG-PET scans should be required, as adding sufficient information (in terms of increased accuracy, and versus pathology, biomarker-based diagnosis or conversion at follow-up) as compared to visual reading alone, to optimize the diagnostic work-up of patients with dementing neurodegenerative disorders.*

Eligibility criteria. Only original full papers published in English on international impacted journals were considered, excluding reviews, management guidelines, abstracts and gray literature. Any sample size was allowed if pathology was the gold standard for diagnosis. Otherwise, 30 subjects, demented patients and/or healthy controls, were requested as the minimum sample sizes.

Literature search. Electronic search strategy, developed and tested with panelists, was performed using predefined strings, grounding on the specific PICO question and including a selection of terms taken from a largely inclusive literature selection, in order to pick all variants for the same keyword. Other details of the literature search are reported in the methodological paper [13].

Data extraction and quality assessment. CF extracted data for this review. The quality of evidence was assessed consensually within the methodology team based on study design, gold/reference standard, FDG-PET image assessment (visual or semi-quantitative methods), risk of bias, index test imprecision, applicability, effect size, and effect inconsistency [13]. A final assessment of relative availability of evidence was formulated, keeping into account evidence availability among all of the 21 PICOs. This ranking was summarized as very poor/lacking, poor, fair or good.

RESULTS

Among the 17 papers identified and screened by the referent panelist (FN), 14 were sent to the methodology team for data extraction and assessment (Figure 1). Two papers [15,16] failed to show semiquantitative results and were not considered. Among the remaining 12, critical outcomes were available in 9 papers (see Table PICO 21; data extraction table available at (https://drive.google.com/file/d/0B0_JB3wzTvbpM0ZmVXZHX2R6c00/view?usp=sharing).

Only three papers performed the direct comparison between visual and automated assessment and quantified the incremental value provided by the latter (see Table PICO 21: Part 1, first row). Perani at al. [5] described a statistically significant change in the level of confidence using the visual assessment of SPM maps as compared to visual assessment of native orthogonal images (2.4 versus 2.07; p=0.003) in a variety of neurodegenerative conditions. Lehman [17] compared visual assessment of native images with visual assessment of 3D-SSP (projection and z-score) maps using GE CortexID® software in patients with MCI or AD dementia. They found the following values for visual assessment of native images and visual assessment of 3D-SSP projection and z-score maps, respectively: sensitivity 0.82 and 0.83, specificity 0.41 and 0.75 (p<0.01), and AUC 0.88 and 0.72 (p=0.017). They also reported a borderline significant increase in confidence with 3D-SSP (p=0.048). Again by comparing visual assessment of native images and of 3D-SSP images, Burdette et al. [18] showed a significant (p=0.043) AUC increase with the automatic tool (from 0.94 to 0.99) in AD patients.

All of the nine selected papers [5,17–23] reported a variety of measures including accuracy, sensitivity, specificity or others (AUC, PPV, NPV, LR+, LR-) of visual and automated FDG-PET assessment in identifying disease-specific hypometabolism. The results are extensively described in Table PICO 21 (Part 1 and Part 2). In general, studies with semi-quantitative and visual reading reported similar ranges of sensitivity (62.3%-96% and 59%-89.6%, respectively) and accuracy (70%-97.5% and 64.8%-89.2%) whereas specificity was greater with automated assessment than with visual reading (range: 84%-99% vs 50%-96%). As expected, the accuracy of visual reading was often found to depend on the skills and experience of the reader[17–20,22]

Four papers[6,21,23,24], based overall on 661 subjects and on good reference standard, calculated both the positive and negative likelihood ratios (LR+ and LR−). All of the analyses performed using semi-quantitative tools achieved LR+ higher than 5 and a LR− ≤0.2 (except PALZ), demonstrating diagnostic utility [25].

With reference to the above mentioned studies, VROI analysis and 3D-SSP were the most frequently used tools, followed by SPM and PALZ. Subjects included in the studies suffered from heterogeneous neurodegenerative disorders (e.g. AD, FTLD, MCI, and Vascular dementia). Risk of bias could be identified in imprecision, use of only reference standard, and applicability of the index test (Table PICO 21, part 1 and 2) as not all memory clinics/nuclear medicine departments may be able to implement these tools.

The level of evidence supporting the clinical utility of semi-automated analysis of FDG-PET was the highest among the PICOs of the entire project, and thus considered as "good". Agreement on the recommendation was achieved on Delphi Round I, 6 panelists supporting the additional use of semi-automated processing to assist visual reading in clinical settings based on its utility both for experienced and non-experienced readers.

DISCUSSION

We assessed the evidence on the utility of computer-aided tools for semi-quantitative assessment of FDG-PET in supporting the nuclear medicine physician to achieve the best interpretation of findings in neurodegenerative disorders.

The main message coming from literature review and expert consensus is that while sensitivity is similar to the one of visual interpretation of native images in this context of differentiation of neurodegenerative disorders, specificity could be better, especially if compared to non-expert readers [21]. They can get confirmation on the intensity and statistical weight of the visually assessed abnormalities, and can thus disambiguate doubtful areas of abnormalities. Indeed, a limited experience on the wide range of findings in normal people could lead to both over- and underestimation of abnormal scans. Another common situation is the presence of symmetric,

moderate hypometabolism in paramedian brain structures, such as the cingulate gyrus, the thalami, or the medial frontal gyrus, that might be missed during visual analysis. Normally these regions have a higher metabolic level compared to the rest of the cortical ribbon, thus early and symmetric hypometabolism can be missed because although the region is hypometabolic, it looks similar to the rest of the cortical ribbon. In fact, the human eye strongly relies on concomitant evaluation of the same structure in the two hemispheres during reporting (i.e., asymmetry evaluation) and thus such findings may be difficult to unveil (Figure 2). Indeed, it has been acknowledged that even the expert readers may benefit from semi-quantitative assessment, mainly by increasing specificity and diagnostic confidence [21].

Other advantages of semi-quantitative assessment are a better standardization of repeated examinations (e.g. for therapy control and follow-up studies) and the option to generate structured reporting (e.g. by reporting regional z-scores) and easier illustration of the findings to the referring physicians, which is especially true for surface rendering approach systems.

A logical consequence of the above reasoning is that automatic semi-quantitative assessment should assist the nuclear medicine physician in interpretation of findings to increase specificity and/or sensitivity and diagnostic confidence. While providing a valuable help, it cannot replace anyway his/her visual analysis, a misunderstanding that may create the false impression that achieving expertise is unnecessary given the availability of an automatic tool. This misunderstanding should be avoided for a variety of reasons. First, the output of the majority of automatic tools is still an image, that is different from native images because it shows voxels of statistically significant lower uptake, normalized on a reference region as compared to a control group, instead of FDG uptake in a color scale. Moreover, stereotactic normalization (warping) is usually performed before statistical comparison changing individual anatomy and thus artificially modifying size and precise location of lesions. In turn, these maps still need to be interpreted, similarly as native images. Thus, the accurate knowledge of hypometabolic patterns of pathological conditions remains the pre-requisite to interpret these statistical maps exactly as for native images.

If a reader does not know, let's say, the typical findings in semantic dementia, the statistical map *per se* cannot teach him/her. Second, those tools producing a score or an index with established cut-offs will never inform about the underlying pathology, but require that the differential diagnosis is restricted *a priori* to no more than two categories, e.g., AD or normal control. This is the case, for instance, for the PALZ score or the HCI which can be used only if the differential diagnosis is between an AD patient and a normal control. Third, with the most popular free software, such as SPM and 3D-SSP, both the reference normalization region and the statistical threshold can be changed which can produce different results for the same scan. For instance, 3D-SSP can use the whole brain, pons, thalami, and cerebellum; SPM uses as a default the whole brain but the normalization region can be changed or even customized to that peculiar comparison, following the Yakushev et al [26] suggestions. These steps must be known and managed with care and require at least a basic knowledge of the software and of statistical rules which is not obvious. Fourth, the standard application of some of these software tools, such as SPM and PALZ, does not correct for atrophy, thus generating maps of significant hypometabolism indifferently for true hypometabolism and for atrophy. The 3D-SSP routine (Neurostat) (https://neurostat.neuro.utah.edu/) has an inherent atrophy correction mechanism implemented as the tool searches for the maximum peak within a predefined distance orthogonally oriented across the cortical ribbon. However, the weight of atrophy may be assessed visually through coregistration of PET with the CT contextually acquired with PET for attenuation correction. Fifth, with SPM there is no standard control group that must be built-up by the user in each center, preferably using the same scanner and the same patient preparation (e.g., eyes open or closed) and reconstruction protocols as the case under analysis. 3D-SSP has a control group that can be modified by users while PALZ, CortexID® and MIMVista® have a fixed control group. Thus, the results of comparisons between a single patient and a control group change depending on the kind and quality of the control group used, besides the reference region chosen for count normalization and the statistical threshold that is chosen determining sensitivity and specificity (see above).

Finally, the results of application of these tools may vary from one another and differences may be relevant, as shown in head-to-head comparisons [9,12]. Figure 3 shows an example of good concordance between three automatic tools and visual reading, whereas Figure 4 highlights substantially different findings depending on the tool used. Thus, these tools are certainly valuable and precious to assist visual reading, but one must be aware of their heterogeneity and intrinsic way of functioning, characterized by different patterns of strengths and limitations. Care should be paid by expert nuclear physicians to train younger personnel in this perspective.

The real world at present tells us that not all nuclear medicine centers use an automatic tool to assist in reporting. Those using them have a large choice among commercial or free methods, often non-CE licensed, and thus there is an almost complete lack of harmonization among centers. This heterogeneity adds to the intrinsic heterogeneity deriving from differences among human readers, despite the teaching efforts of the EANM as well as of the National nuclear medicine societies. Finally, scanner manufacturers add in turn their part of inhomogeneity by implementing different tools in the working stations.

As a future perspective, the scientific societies in the field could promote studies that compare the performance of different tools and guide centers to implement the most accurate, cheap, and easy-to-use, hopefully supported in this action by manufacturers. These performances should always be compared with the reading of a group of experts.

Acknowledgements

Compliance with Ethical Standards.

Conflict of Interest:

Flavio Nobili: received personal fees and non-financial support from GE Healthcare, non-financial support from Eli-Lilly and grants from Chiesi Farmaceutici

Cristina Festari: declares that she has no conflict of interest

Daniele Altomare: was the recipient of the grant allocated by the European Academy of Neurology (EAN) for data extraction and evidence assessment for the present project.

Federica Agosta: is Section Editor of NeuroImage: Clinical; has received speaker fees from Biogen Idec, Novartis, and Excellence in Medical Education; and receives or has received research supports from the Italian Ministry of Health, AriSLA (Fondazione Italiana di Ricerca per la SLA), and the European Research Council. She received personal fees from Elsevier INC.

Stefania Orini: declares that she has no conflict of interest

Koen Van Laere: declares that she has no conflict of interest

Javier Arbizu: received grants from Eli-Lilly & Co, Piramal and GE Healthcare

Femke Bouwman: declares that she has no conflict of interest

Peter Nestor: declares that he has no conflict of interest

Alexander Drzezga: received grants and non-financial support from Eli-Lilly & Co, Siemens and GE Healthcare; he also received non-financial support from Piramal

Ethical approval: This article does not contain any new studies with human participants or animals performed by any of the authors. The human studies discussed herein came exclusively from previously published research articles.

Informed consent: not applicable, this is a review article. Informed consent statement is declared in each of the revised paper.

References

1. Nobili F, Arbizu J, Bouwman F, Drzezga A, Filippi M, Nestor P, et al. EAN-EANM recommendations for the use of brain 18F-Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) in neurodegenerative cognitive impairment and dementia: Delphi consensus. Eur J Neurol. (submitted)

2. Varrone A, Asenbaum S, Vander Borght T, Booij J, Nobili F, Någren K, et al. EANM procedure guidelines for PET brain imaging using [18F]FDG, version 2. Eur J Nucl Med Mol Imaging. 2009;36:2103–10.

3. Friston K, Holmes A, Worsley K, Poline J, Frith C, Frackowiak R. Statistical parametric maps in functional imaging: A general linear approach. Hum Brain Mapp. 1994;2:189–210.

4. Signorini M, Paulesu E, Friston K, Perani D, Colleluori A, Lucignani G, et al. Rapid Assessment of Regional Cerebral Metabolic Abnormalities in Single Subjects with Quantitative and Nonquantitative [18F]FDG PET: A Clinical Validation of Statistical Parametric Mapping. Neuroimage.1999;9:63–80.

5. Perani D, Della Rosa PA, Cerami C, Gallivanone F, Fallanca F, Vanoli EG, et al. Validation of an optimized SPM procedure for FDG-PET in dementia diagnosis in a clinical setting. NeuroImage Clin. 2014;6:445–54.

6. Della Rosa PA, Cerami C, Gallivanone F, Prestia A, Caroli A, Castiglioni I, et al. A Standardized [18F]-FDG-PET Template for Spatial Normalization in Statistical Parametric Mapping of Dementia. Neuroinformatics. 2014;12:575–93.

7. Minoshima S, Frey KA, Koeppe RA, Foster NL, Kuhl DE. A diagnostic approach in Alzheimer's disease using three-dimensional stereotactic surface projections of fluorine-18-FDG PET J Nucl Med. 1995;36:1238-48

8. Herholz K, Salmon E, Perani D, Baron JC, Holthoff V, Frölich L, et al. Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG PET. Neuroimage. 2002;17:302–16.

9. Caroli A, Prestia A, Chen K, Ayutyanont N, Landau SM, Madison CM, et al. Summary Metrics to Assess Alzheimer Disease-Related Hypometabolic Pattern with 18F-FDG PET: Head-to-Head Comparison. J Nucl Med. 2012;53:592–600.

10. Chen K, Ayutyanont N, Langbaum JBS, Fleisher AS, Reschke C, Lee W, et al. Characterizing Alzheimer's disease using a hypometabolic convergence index. Neuroimage. 2011;56:52–60.

11. Landau SM, Harvey D, Madison CM, Koeppe RA, Reiman EM, Foster NL, et al. Associations between cognitive, functional, and FDG-PET measures of decline in AD and MCI. Neurobiol Aging. 2011;32:1207–18.

12. Pagani M, De Carli F, Morbelli S, Öberg J, Chincarini A, Frisoni GB, et al. Volume of interest-based [18F]fluorodeoxyglucose PET discriminates MCI converting to Alzheimer's disease from healthy controls. A European Alzheimer's Disease Consortium (EADC) study. NeuroImage Clin. 2015;7:34–42.

13. Boccardi M, Festari C, Altomare D, Gandolfo F, Orini S, Nobili F, et al. Assessing accuracy diagnostic FDG-PET studies to define clinical use for dementia diagnosis. EJNMMI.

14. Leone MA, Brainin M, Boon P, Pugliatti M, Keindl M, Bassetti CL. Guidance for the preparation of neurological management guidelines by EFNS scientific task forces - revised recommendations 2012. Eur J Neurol. 2013;20:410–9.

15. Patterson JC, Lilien DL, Takalkar A, Kelley RE, Minagar A. Potential value of quantitative analysis of cerebral PET in early cognitive decline. Am J Alzheimers Dis Other Demen. 2009;23:586–92.

16. Von Borczyskowski D, Wilke F, Martin B, Brenner W, Clausen M, Mester J, et al. Evaluation of a new expert system for fully automated detection of the Alzheimer's dementia pattern in FDG PET. J Nucl Med. 2006;27:739–43.

17. Lehman VT, Carter RE, Claassen DO, Murphy RC, Lowe V, Petersen RC, et al. Visual assessment versus quantitative three-dimensional stereotactic surface projection fluorodeoxyglucose positron emission tomography for detection of mild cognitive impairment and Alzheimer disease.

Clin Nucl Med. 2012;37:721–6.

18. Burdette JH, Minoshima S, Vander Borght T, Tran DD, Kuhl DE. Alzheimer disease: improved visual interpretation of PET images by using three-dimensional stereotaxic surface projections. Radiology. 1996;198:837–43.

19. Ishii K, Kono AK, Sasaki H, Miyamoto N, Fukuda T, Sakamoto S, et al. Fully automatic diagnostic system for early- and late-onset mild Alzheimer's disease using FDG PET and 3D-SSP. Eur J Nucl Med Mol Imaging. 2006;33:575–83.

20. Kono AK, Ishii K, Sofue K, Miyamoto N, Sakamoto S, Mori E. Fully automatic differential diagnosis system for dementia with Lewy bodies and Alzheimer's disease using FDG-PET and 3D-SSP. Eur J Nucl Med Mol Imaging. 2007;34:1490–7.

21. Morbelli S, Brugnolo A, Bossert I, Buschiazzo A, Frisoni GB, Galluzzi S, et al. Visual Versus semi-quantitative analysis of 18F-FDG-PET in Amnestic MCI: An European Alzheimer's Disease Consortium (EADC) project. J Alzheimer's Dis. 2015;44:815–26.

22. Ng S, Villemagne VL, Berlangieri S, Lee S-T, Cherk M, Gong SJ, et al. Visual Assessment Versus Quantitative Assessment of 11C-PIB PET and 18F-FDG PET for Detection of Alzheimer's Disease. J Nucl Med. 2007;48:547–52.

23. Rabinovici GD, Rosen HJ, Alkalay A, Kornak J, Furst AJ, Agarwal N, et al. Amyloid vs FDG-PET in the differential diagnosis of AD and FTLD. Neurology. 2011;77:2034–42.

24. Foster NL, Heidebrink JL, Clark CM, Jagust WJ, Arnold SE, Barbas NR, et al. FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease. Brain. 2007;130:2616–35.

25. Frisoni GB, Bocchetta M, Chételat G, Rabinovici GD, De Leon MJ, Kaye J, et al. Imaging markers for Alzheimer disease: Which vs how. Neurology. 2013. p. 487–500.

26. Yakushev I, Hammers A, Fellgiebel A, Schmidtmann I, Scheurich A, Buchholz HG, et al. SPM-based count normalization provides excellent discrimination of mild Alzheimer's disease and amnestic mild cognitive impairment from healthy aging. Neuroimage. 2009;44:43–50.

27. Moher D1, Liberati A, Tetzlaff J ADPG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 2009;62:1006–12.

FIGURE LEGENDS

Figure 1. PRISMA flowchart of selected papers for PICO 21 regarding requirement of semi-automated assessment (adapted from Moher et al. [27]).

FIGURE 2. FDG-PET native images (left) of a 78 years old woman with MCI due to AD. MMSE score = 23/30; CDR= 0.5. The bilateral, significant hypometabolism of posterior cingulate cortex is easily appreciated with an automatic tool in comparison with a normal control group, corrected for age, uncorrected for atrophy (right; PALZ, clusters of significant hypometabolism in red; [8] while it can be difficult to be detected on native images.

FIGURE 3. FDG-PET native images (a, e) of a 53 years old man with MCI due to AD. MMSE score = 28/30; CDR= 0.5. On visual analysis a moderate hypometabolism of right lateral temporal and parietal cortex was found together with bilateral hypometabolism of posterior cingulate cortex. These findings were confirmed with automatic analysis by PALZ (b, f), 3D-SSP (c: lateral projections; g: medial projections; clusters of significant hypometabolism in green-yellow; [7] which also showed clusters in anterior cingulate cortex and medial temporal lobe, not found with either PALZ or visual analysis, and SPM[4] using a height threshold of p<0.01 (h) while with the more restrictive threshold of p<0.001 only the posterior cingulate hypometabolism was highlighted (d). Other details as in Figure 2.

FIGURE 4. FDG-PET native images (a, e) of a 56 years old healthy woman with doubtful MCI and a positive family history for AD. MMSE score = 30/30; CDR= 0. On visual analysis a mild-to-moderate hypometabolism of left lateral temporal and parietal cortex was observed. These findings were confirmed with automatic analysis by PALZ (b, f), also showing significant clusters in bilateral orbitofrontal cortex while 3D-SSP (c: lateral projections; g: medial projections) essentially

failed to show left temporo-parietal hypometabolism and disclosed instead some medial frontal and temporal hypometabolic clusters. SPM did not show any significant hypometabolic cluster with a height threshold of p<0.001 or p<0.01 (d) and only a very permissive threshold of p<0.05 highlighted clusters of significant hypometabolism in left lateral frontal, temporal, and parietal cortex (h). Other details as in Figure 2-3.

Table 1. PICO 21 (PART 1). Table reports the quality of evidence for each critical outcome.

Critical outcomes	N. of papers	Sample size	Gold/ reference standard	Risk of bias	Index test imprecision	Applicability	FDG-PET assessment	Effect range (CI)	Effect assessment	Effect inconsistency	Outcome quality
PICO 21: Incremental value of automated assessment of FDG-PET compared to visual reading											
Incremental value indices	3	156 Patients 157 HC	2 Diagnosis at follow-up 1 Clinical diagnosis	Serious	Serious	Not serious	Visual + Semi-quantitative (1 SPM-Maps, 2 3D-SSP)	**Study 1 (Visual vs SPM-Maps).** - Level of confidence 2.07 vs 2.4, p=0.003. **Study 2 (Visual vs 3D-SSP).** - Sensitivity: 83% (CI: 66-94%) vs 82% (CI: 62-92%), p=1.0. - Specificity: 41% (CI: 20-61%) vs 75% (CI: 52-90%), p<0.01. - AUC: 72% (CI: 55-83%) vs 88% (CI: 76-95%), p=0.017. - Mean increase in confidence rating = 0.7 (CI: 0.01-1.3), p=0.048. **Study 3 (Visual vs 3D-SSP).** - AUC: 94% (SD 0.03) vs 0.99 (SD 0.01), p = 0.043.	MODERATE	NA	LOW
Sensitivity	6	479 Patients 126 HC	1 Pathology 1 Biomarker-based diagnosis 2 Diagnosis at follow-up 2 Clinical diagnosis	Not serious	Serious	Not serious	Visual	59% (range: 43-71%)– 89.6% (CI 80-95%)	MODERATE	Serious	MODERATE
							Semi-quantitative (2 ROI, 2 3D-SSP, 1 SPM, 1 PALZ)	62.3% (CI 50-73%) – 96% (CI NA)	MODERATE	Serious	MODERATE
Specificity	6	479 Patients 126 HC	1 Pathology 1 Biomarker-based diagnosis 2 Diagnosis at follow-up 2 Clinical diagnosis	Not serious	Serious	Not serious	Visual	50% (CI NA) – 96% (range: 92-100%)	HIGH	Very serious	LOW
							Semi-quantitative (2 ROI, 2 3D-SSP, 1 SPM, 1 PALZ)	84% (CI NA) – 99% (SD 0.02)	HIGH	Not serious	HIGH
Accuracy	7	459 Patients 237 HC	1 Pathology 1 Biomarker-based diagnosis 2 Diagnosis at follow-up 3 Clinical diagnosis	Serious	Not serious	Not serious	Visual	64.8% (CI: 51-77%) – 89.2% (CI: 84–93%)	MODERATE	Serious	MODERATE
							Semi-quantitative (3 ROI, 3 3D-SSP, 1 PALZ)	70% (CI: 53-84%) – 97.5% (CI: 91-100%)	HIGH	Serious	MODERATE

Table 1. PICO 21 (PART 2).

PICO 21: Incremental value of automated assessment of FDG-PET compared to visual reading											
Critical outcomes	**N. of papers**	**Sample size**	**Gold/ reference standard**	**Risk of bias**	**Index test imprecision**	**Applicability**	**FDG-PET assessment**	**Effect range (CI)**	**Effect assessment**	**Effect inconsistency**	**Outcome quality**
AUC	3	155 Patients 142 HC	1 Diagnosis at follow-up 2 Clinical diagnosis	Serious	Serious	Not serious	Visual	50% (CI NA) ; 87.8 (CI NA)	MODERATE	Serious	LOW
							Semi-quantitative (1 ROI, 1 SPM, 1 3D-SSP)	67 (CI NA);96.7 (CI NA)	HIGH	Serious	LOW
PPV	3	294 Patients 167 HC	1 Pathology 1 Biomarker-based diagnosis 1 Diagnosis at follow-up	Not serious	Serious	Not serious	Visual	68% (range: 50-88%) – 87.5% (CI NA)	HIGH	Not serious	MODERATE
							Semi-quantitative (1 ROI, 1 3D-SSP, 1 PALZ)	84.2% (CI: 72–92%) – 98% (CI: 88–100%)	HIGH	Not serious	MODERATE
NPV	3	294 Patients 167 HC	1 Pathology 1 Biomarker-based diagnosis 1 Diagnosis at follow-up	Not serious	Serious	Not serious	Visual	72% (CI NA) – 92.4% (CI: 85-96%)	HIGH	Serious	MODERATE
							Semi-quantitative (1 ROI, 1 3D-SSP, 1 PALZ)	71% (CI: 58-93%) – 89% (range: 85-92%)	MODERATE	Not serious	MODERATE
LR+	4	382 Patients 279 HC	1 Pathology 1 Biomarker-based diagnosis 2 Diagnosis at follow-up	Not serious	Serious	Serious	Visual	1.55 (CI NA) – 14.8 (CI: 10.7-∞)	MODERATE	Serious	LOW
							Semi-quantitative (1 ROI, 1 SPM, 1 3D-SSP, 1 PALZ)	6.08 (CI NA) – 36.5 (CI: 21.3-∞)	HIGH	Very serious	LOW
LR-	4	382 Patients 279 HC	1 Pathology 1 Biomarker-based diagnosis 2 Diagnosis at follow-up	Not serious	Serious	Serious	Visual	0.12 (CI: 0.06-0.23) – 0.45 (CI NA)	MODERATE	Serious	LOW
							Semi-quantitative (1 ROI, 1 SPM, 1 3D-SSP, 1 PALZ)	0.03 (CI: 0.0-0.5) – 0.41 (CI: 0.31-0.55)	MODERATE	Very serious	LOW
RELATIVE AVAILABILITY OF EVIDENCE: STRONG											

Risk of bias: assessment of the study design and other methodological features (e.g., patient selection, clinical diagnostic criteria used). Index test methods: assessment of index test methodology (e.g., technical details, image analysis methods and statistical analysis). Applicability: representativeness of the studied population and index test reproducibility in clinical practice (semi-quantitative methods correspond to 'serious' indirectness, visual + semi-quantitative methods correspond to 'not serious' indirectness, due to partial implementation of quantitation in clinical practice). Effect: lowest and highest values for each critical outcome; when more values were obtained for the same outcome, the highest was reported. Effect assessment: 51-70% low, 71-80% moderate, 81-100% high. Effect inconsistency: 'Not serious' if lowest and highest values difference was 0-20, 'serious' 21-40, 'very serious' >40. Outcome quality: summary of evidence as from all columns.

a

b

c

d

e

f

g

h

a

b

c

d

e

f

g

h