Assessing accuracy diagnostic FDG-PET studies to define clinical use for dementia diagnosis

Marina Boccardi[1,2*], Cristina Festari[2,3], Daniele Altomare[2,3], Federica Gandolfo[4], Stefania Orini[4], Emiliano Albanese[5] Flavio Nobili[6], Giovanni B Frisoni[1,2,7], *for the EANM-EAN Task Force for the Prescription of FDG-PET for Dementing Neurodegenerative Disorders*

Collaborators: Federica Agosta[8], Javier Arbizu[9], Femke Bouwman[10], Alexander Drzezga[11], Peter Nestor[12], Zuzana Walker[13].

[1]LANVIE – Laboratory of Neuroimaging of Aging of Aging, University of Geneva, Geneva, Switzerland.

[2]LANE – Laboratory of Alzheimer Neuroimaging & Epidemiology, IRCCS S. Giovanni di Dio, Fatebenefratelli, Brescia, Italy.

[3]Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy.

[4] Alzheimer Operative Unit; IRCCS Centro S. Giovanni di Dio, Fatebenefratelli, The National Center for Research and Care of Alzheimer's and Mental Diseases, Brescia, Italy;

[5]WHO Collaborating Center, Department of Psychiatry, University of Geneva, Geneva, Switzerland.

[6]DINOGMI – Department of Neuroscience, University of Genoa and IRCCS AOU San Martino-IST Genoa, Italy.

[7]HUG Hopitaux Universitaires de Genève, Geneva, Switzerland.

[8] Neuroimaging Research Unit, Institute of Experimental Neurology, Division of Neuroscience, San Raffaele Scientific Institute, Vita-Salute San Raffaele University, Milan, Italy.

[9] Department of Nuclear Medicine. Clinica Universidad de Navarra. University of Navarra. Pamplona, Spain.

[10] Department of Neurology & Alzheimer Center, Amsterdam Neuroscience, VU University Medical Center, Amsterdam, the Netherlands.

[11] Department of Nuclear Medicine, University Hospital of Cologne, University of Cologne and German Center for Neurodegenerative Diseases (DZNE), Germany.

[12] German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany; Queensland Brain Institute, University of Queensland and at the Mater Hospital Brisbane.

[13] University College London, Division of Psychiatry & Essex Partnership University NHS Foundation Trust, UK.

*Marina Boccardi

LANVIE (Laboratoire de Neuroimagerie du Vieillissement), Dept of Psychiatry

University of Geneva

Chemin du Petit-Bel-Air, 2

1225, Chene-Bourg, Genève, Suisse

e-mail: marina.boccardi@unige.ch

Tel.: 0041.(0)22.3055764, Fax.: 0041.(0)22.3054719

Abstract

Background: FDG-PET is frequently used as a marker of synaptic damage to diagnose dementing neurodegenerative disorders. We aimed to adapt the items of evidence quality to FDG-PET diagnostic studies, and assess the evidence available in current literature and assist Delphi decisions for European recommendations for clinical use.

Methods: based on acknowledged methodological guidance, we defined the domains, specific to FDG-PET, required to assess the quality of evidence in 21 literature searches addressing as many PICO questions. We ranked findings for each PICO and fed experts taking Delphi decisions for recommending clinical use.

Results: Among the 1435 retrieved studies, most lacked validated measures of test performance, adequate gold standard, and head-to-head comparison of FDG-PET and clinical diagnosis, and only 58 entered detailed assessment. Only 2 studies assessed the accuracy of the comparator (clinical diagnosis) versus any kind of gold-/reference-standard. As to the index-test (FDG-PET-based diagnosis), an independent gold-standard was available in 24% of the examined papers; 38% used an acceptable reference-standard (clinical follow-up); 38% compared FDG-PET-based diagnosis only to baseline clinical diagnosis. These methodological limitations did not allow to derive recommendations from evidence.

Discussion: An incremental diagnostic value of FDG-PET versus clinical diagnosis or lack thereof cannot be derived from current literature. Many of the observed limitations may easily be overcome, and we outlined them as research priorities to improve the quality of current evidence. Such improvement is necessary to outline evidence-based guidelines. The available data were anyway provided to expert clinicians who defined interim recommendations.

Keywords: FDG-PET; Positron emission tomography, Fluorodeoxyglucose metabolism, 18F-FDG PET; evidence assessment; quality of evidence; recommendations; validation; biomarker; Alzheimer; dementia; diagnosis.

1. Introduction

Notwithstanding their still limited validation (1), neuroimaging biomarkers that assess neurodegeneration are crucial in the diagnosis of dementing disorders. They help to unveil whether observed cognitive impairment is associated to a neurodegenerative condition, to be further identified with the use of pathophysiological biomarkers, that are however not available for all diseases to date (2). Moreover, neuroimaging biomarkers can inform about the stage of the disorder, neuronal damage constituting the latest biological event in their long pathophysiological course (3), and the one better corresponding to, and predicting the, severity of symptoms (2).

Among neuroimaging biomarkers, FDG-PET is used very frequently to ascertain the presence of neurodegeneration, even at early symptomatic stages, for its sensitivity to synaptic dysfunction, an event that precedes neuronal death. Being able to detect hypofunction in the first cerebral regions targeted by the disorder, the pattern of hypometabolism is also used to address diagnosis, based on current knowledge of the circuits typically targeted in the different conditions. Unfortunately, though, our knowledge of the pathophysiology of these disorders is still limited, and correspondence of clinical syndromes, hypometabolic systems, and underlying pathologies imperfect.

Given this context, and considering that FDG-PET is used frequently in the clinical diagnosis of neurodegenerative conditions in the lack of evidence-based guidelines, we explored the quality of available FDG-PET diagnostic studies, to see whether they provide the evidence required to assess its clinical validity (4), and outline usage guidelines (5, 6). We performed such assessment for a wide range of clinical scenarios, to assist decisions for the joint European Association of Nuclear Medicine (EANM) and European Academy of Neurology (EAN) recommendations aimed to guide clinicians in the diagnostic use of FDG-PET (7). The aim of this paper is to describe how we adapted the domains of evidence quality assessment to the field of FDG-PET-based diagnosis within current methodological frameworks specific to diagnostic studies (5, 8-12), and provide an overall report of the quality of current studies. Based on our results, we also outlined the most urgent and feasible improvements that current research may and should implement, for a substantial advancement of the validation of this particularly useful biomarker.

2. Methods

2.1 Project structure

By initiative of the EANM, and in association with the EAN, seven experts in FDG-PET and neurodegenerative disorders have been nominated from the two societies (7). They outlined 21 Population-Intervention-Comparison-Outcome (PICO) questions addressing urgent clinical issues requiring guidance for FDG-PET use in the setting of memory clinics (Table 1), and performed PICO-structured searches as appropriate (Table 2). Papers were screened to include those reporting the comparison of interest and quantitatively assessing the index-test performance, and were assessed as outlined in the following sections. The seven panelists were appointed to produce recommendations taking into consideration the so assessed incremental value of FDG-PET, as added to clinical-neuropsychological examination, for the diagnosis and management of patients with dementing neurodegenerative disorders of different types. Consensus recommendations have been produced with a Delphi procedure based on the expertise of panelists, informed about the availability and quality of evidence(7).

Table 1. *Questions requiring the definition of recommendations for the clinical use of FDG-PET. The questions were formulated based on the PICO structure (Population, Intervention, Comparison, Outcome) to explicitly address all of the relevant methodological parameters for evidence assessment. The role of FDG-PET consists of supporting diagnosis as an add-on to traditional clinical and neuropsychological assessment. The assessment of available literature is aimed to explore whether the exam has sufficient incremental diagnostic to justify such use.*

PICO	Should FDG-PET be performed, as adding diagnostic value (in terms of increased accuracy, and versus pathology or biomarker-based diagnosis or conversion at follow-up) as compared to standard clinical/neuropsychological assessment alone, to:
1	detect Alzheimer's disease in patients with persistent MCI of uncertain origin?
2	detect fronto-temporal lobar degeneration in patients with persistent MCI of uncertain origin?
3	detect prodromal Lewy bodies dementia in patients with persistent MCI of uncertain origin?
4	pick early signs of neurodegeneration in patients with subjective cognitive decline?
5	pick early signs of neurodegeneration in patients with asymptomatics at risk for AD?
6	detect early signs of neurodegeneration in asymptomatic carriers of AD mutation?
7	differentiate among main forms of dementia in patients with dementia and either atypical presentation or atypical course?
8	differentiate between Alzheimer Disease and dementia with Lewy Bodies?
9	differentiate Alzheimer disease from fronto-temporal lobar degeneration?
10	differentiate between dementia with Lewy Bodies and fronto-temporal lobar degeneration?
11	differentiate between Alzheimer disease and vascular dementia?
12	identify brain dysfunction related to cognitive deterioration in patients with PD and cognitive impairment?
13	discriminate PSP from Parkinson's disease?

14	discriminate pseudodementia?
15	differentiate the underlying pathological process in patients with corticobasal syndrome?
16	obtain indirect information on the molecular pathologies in patients with primary aphasias?
17	confirm a clinical suspicion of ALS in patients with or without cognitive impairment?
18	detect brain dysfunction related to cognitive deterioration in patients with ALS?
19	pick early signs of neurodegeneration in patients with a genetic risk of Huntington disease?
20	discriminate frontal-lobe hypometabolism involved in cognitive deterioration in patients with Huntington disease?
21	Should automated assessment of FDG-PET scans 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 as taken alone, to optimize the diagnostic work-up of patients with dementing neurodegenerative disorders?

2.2 Terminology

Throughout the whole project, we distinguished between clinical syndromes and pathophysiological disorders, as started by the advent of biomarkers for AD (13, 14). The separation of the concepts of clinical stage, syndrome and pathophysiology (15) is potentially appropriate also for other disorders, like FTLD or DLB, for which it is being considered (16), but is not yet fully outlined. Although sometimes potentially controversial, such approach was required to answer PICO questions like those about "detecting DLB (*dementia* with Lewy Bodies) *in MCI patients*".

Although this paper is not meant as providing an exhaustive definition of methodological assessment for diagnostic studies, we define below some key terms to the advantage of non specialized readers; more information should be found in the methodological literature on evidence assessment (http://methods.cochrane.org/sdt/handbook-dta-reviews) (8, 10-12, 17, 18).

Index test: the test which diagnostic performance is assessed (FDG-PET in our case);

Comparator: the test versus which the index test is compared, in order to assess its incremental value; in our case, the comparator is traditional clinical and neuropsychological examination. In the EANM-EAN initiative, FDG-PET is meant as an add-on to the traditional clinical procedure. However, diagnostic studies designed to compare the performance of index-test and comparator are the tool allowing to make decisions also for add-on tests.

Gold-standard: the most accurate independent test demonstrating absence or presence of the target pathology (e.g, autopsy confirmation).

Reference-standard: the best available independent diagnostic confirmation under reasonable conditions (e.g., clinical diagnosis at follow-up).

Critical outcome: critical outcomes are the quantitative measures that allow objective assessment of literature quality. In the case of diagnostic studies, the appropriate critical outcomes are the

validated measures of test performance (sensitivity, specificity, accuracy, AUC, PPV, NPV, positive and negative likelihood ratios) quantifying the ability of FDG-PET to detect the appropriate clinical diagnosis, as assessed with comparison with the appropriate gold- or reference-standard.
Head-to-head comparison: comparative analysis of the performance of the index test and the comparator, both being assessed versus the same independent gold- or reference-standard.
Incremental diagnostic value: difference in the amount of information allowed by an examination added to a diagnostic procedure. The computation of such difference requires that both the traditional diagnostic procedure and that the added examination are tested versus the same gold- or reference-standard (see head-to-head comparison).

2.3. *Literature searches and eligibility*

The electronic search strategy was performed using strings composed of an FDG-PET diagnostic component, common to all PICOs, and of parts specific to each PICO question. Terms selection was largely inclusive to pick variants (Table 2). Syntaxes were adapted to the following databases: Embase, Pubmed, Google Scholar and CrossReference. Papers including the comparison of interest and the minimum sample size published up to November 2015 were included in the assessment. No minimum sample size was set whenever pathology-based gold-standard was available; otherwise, it was set by the referent panelist, based on the frequency of the disorder and the sample sizes normally available in the literature (Table 2). Studies were first hand-searched by the referent panelist, who could include additional papers, as from personal knowledge or tracking from papers references. Panelists also made a first screening based on abstracts. The full text of these potentially eligible studies has then been independently assessed for eligibility by the methodology team.

Table 2. PICO-specific strings and minimum sample size used for the 21 literature reviews. Minimum sample size refers to the minimum number of subjects required for including papers in the assessment procedure. Each string was used in combination with the common FDG-PET string: *((("Positron emission tomography"[Title] OR "Cerebral positron emission tomography" [Title] OR "PET" [Title]) AND ("Fluorodeoxyglucose" [Title] OR "FDG"[Title] OR "glucose metabolism"[Title] OR "Cerebral metabolic rate of glucose"[Title] OR "Metabolism"[Title] OR "metabolic activity"[Title] OR "metabolic networks"[Title] OR "Hypometabolism"[Title])) OR ("FDG PET"[Title] OR "FDG-PET"[Title] OR "18F-FDG PET"[Title])) NOT review*

PICO	PICO-specific string	Minimum sample size
1	“MCI” OR “Mild cognitive impairment” OR “prodromal” OR conver*) AND “Alzheimer”	15
2	'mci' OR 'mild cognitive impairment'/exp OR 'mild cognitive impairment' OR 'prodromal' AND ('differential diagnosis'/exp OR 'differential diagnosis') AND ('ftd' OR 'ftld'/exp OR 'ftld' OR 'frontotemporal' OR 'fronto-temporal') AND ('positron emission tomography'/exp OR 'positron emission tomography' OR 'cerebral positron emission tomography' OR 'pet' AND ('fluorodeoxyglucose'/exp OR 'fluorodeoxyglucose' OR 'fdg' OR 'glucose metabolism'/exp OR 'glucose metabolism' OR 'cerebral metabolic rate of glucose' OR 'metabolism'/exp OR 'metabolism' OR 'metabolic activity' OR 'metabolic networks' OR 'hypometabolism') OR 'fdg pet' OR 'fdg-pet' OR '18f-fdg pet') NOT ('review'/exp OR review)	15
3	First string((“MCI”[title] OR “Mild cognitive impairment” [title] OR “prodrom*”[title] OR conver*)[title] AND (“Lewy”[title] OR “DLB”[title] OR “LBD”[title]); Second string: (“Lewy”[Title/abstract] OR “DLB”[Title/abstract] OR “LBD”[Title/abstract]) and (MCI[title/abstract] OR “cognitive impairment”[Title/abstract] OR prodrom*[Title/abstract] OR convers*[title/abstract]) AND (“Fluorodeoxyglucose”[Title] OR “Fluoro-deoxyglucose”[Title] OR “FDG”[Title] OR metabol*[Title] OR Hypometabol*[Title] OR “FDG PET”[Title] OR “FDG-PET”[Title] OR “18F-FDG PET”[Title] OR “18F-FDG-PET”[Title])	15
4	(((subjective[All Fields] AND ("cognition disorders"[MeSH Terms] OR ("cognition"[All Fields] AND "disorders"[All Fields]) OR "cognition disorders"[All Fields] OR ("cognitive"[All Fields] AND "impairment"[All Fields]) OR "cognitive impairment"[All Fields])) AND (English[lang] AND medline[sb])) OR ((subjective[All Fields] AND ("memory"[MeSH Terms] OR "memory"[All Fields]) AND complaints[All Fields]) AND (English[lang] AND medline[sb])) AND (English[lang] AND medline[sb])) AND ((fdg[All Fields] AND pet[All Fields]) AND (English[lang] AND medline[sb])) AND (English[lang] AND medline[sb]) AND (Journal Article[ptyp] AND Humans[Mesh] AND English[lang])	15
5	“Alzheimer” AND (“preclinical” OR “asymptomatic” OR APOE OR amyloid)	any
6	“Alzheimer” AND (“familial” OR “genetic” OR “dominant” OR “ADAD” OR “autosomal” OR “presenilin” OR “PSEN1” OR “PSEN2” OR “APP”) AND (“preclinical” OR “asymptomatic”)	5
7	(“Alzheimer” OR “dementia”) AND (“atypical” OR “focal” OR “posterior” OR “logopenic” OR “frontal variant”)	5
8	“Alzheimer”[Title] AND (“Lewy”[Title] OR “DLB”[Title] OR “LBD”[Title])	any
9	“Alzheimer” AND (“FTLD” OR “FTD” OR “frontotemporal” OR “fronto-temporal”) AND “differential diagnosis”	20
10	(“Lewy”[Title] OR “DLB”[Title] OR “LBD”[Title]) AND (“FTLD”[Title] OR “FTD”[Title] OR “frontotemporal”[Title] OR “fronto-temporal”[Title])	any
11	“Alzheimer” AND (“Vascular” OR “subcortical” OR “small vessels disease”) AND “differential diagnosis”	any
12	“Parkins*[title] AND (cognit*[title] OR “decline”[title] OR “deterioration”[title] OR “impairment”[title] OR “dementia”[title] OR “MCI”[title] OR “PDD”[title] OR “PD-MCI”[title]) AND (“Fluorodeoxyglucose”[Title] OR “Fluoro-deoxyglucose”[Title] OR “FDG”[Title] OR metabol*[Title] OR Hypometabol*[Title] OR “FDG PET”[Title] OR “FDG-PET”[Title] OR “18F-FDG PET”[Title] OR “18F-FDG-PET”[Title])	20
13	(“Progressive supra-nuclear palsy” OR “Progressive supranuclear palsy” OR “PSP”) AND “Parkinson” AND “differential diagnosis”	5
14	((“depression” AND neurodeg*) AND (“disease” OR “disorder”)) OR (“pseudo-dementia” OR “depressive pseudo-dementia”) AND “differential diagnosis”	any
15	(“corticobasal” OR “cortico-basal”) AND (“degeneration” OR “neurodegeneration” OR “disease”)	5

16	"Primary progressive aphasia" AND ("logopenic" OR "progressive nonfluent aphasia" OR "progressive non-fluent aphasia" OR "semantic" OR "agrammatic") AND "differential diagnosis"	5
17	("amyotrophic lateral sclerosis" OR "motor neuron disease") AND "diagnosis"	10
18	("amyotrophic lateral sclerosis" OR "motor neuron disease") AND "frontal" AND ("dysfunction" OR "symptoms" OR "syndrome")	5
19	"Huntington" AND ("preclinical" OR "asymptomatic") AND "diagnosis"	5
20	"Huntington" AND "frontal" AND ("dysfunction" OR "symptoms" OR "syndrome")	5
21	"assessment" AND ("visual reading" OR "visual assessment" OR "visual evaluation" OR "automated" OR "quantitative" OR "computer-aided") AND ("cerebral" OR "brain" OR "dementia" OR "neuro*") AND diagnosis AND ("added value" OR "incremental value")	30

2.4 Data extraction

Data extraction was set and performed by the methodology team, including researchers with experience in consensus procedures and methodology (MB, CF and DA) and in clinical practice (SO, FG). We extracted information relating to a large set of variables (see supplemental material at https://drive.google.com/open?id=0B0_JB3wzTvbpVFYtUGxHdGZWYmc), consistent with currently accepted guidance (5, 8, 12) . The extracted data included:

- Study characteristics: author, year of publication, citation rate, study design, sample size, duration of follow-up;
- Population features: demographics, clinical and neuropsychological features of the studied samples (e.g., sample size, age, gender, clinical diagnosis, clinical criteria, MMSE score, CDR score, duration of illness, patient recruitment and accounting; time to conversion or follow-up if pertinent);
- Index test features: scanner technical details (those older than 2005 being considered as possible cause of inconsistencies(19)), scan reading and statistical analysis;
- Reference-/gold-standard features: diagnostic criteria, use of biomarkers;
- Critical outcomes: sensitivity, specificity, accuracy, positive/negative predictive value (PPV/NPV), area under the curve (AUC), or positive/negative likelihood ratios (LR+/LR-); other critical outcomes if applicable.

Data were extracted by a single reviewer for each PICO. When possible, we computed missing values and confidence intervals for the critical outcomes. In case of inconsistencies (20) we recomputed values based on the data provided in the paper.

2.5 Assessment of the quality of evidence

The assessment was based on study design, scan reading procedure, risk of bias, index test imprecision, applicability, effect size, total number of subjects, and effect inconsistency (Table 3) (10-12). Reviewers assessed the quality of evidence of individual studies independently. Then, the global assessment of each outcome (i.e., each of the different measures of test performance, such as accuracy, AUC, PPV, etc) across studies (18, 21) was proposed by the data extractor and then discussed and fine-tuned consensually within the methodology working group, keeping into account the quality of the individual source studies.

Based on the resulting assessment, the quality of evidence was then ranked within the 21 PICOs. More precisely, PICOs lacking critical outcomes entirely were put at the lowest level, while those with soundest methodology, numerous studies, large total number of included subjects, and large and consistent effect size and were graded best. The other PICOs were ranked in between. In this way, we provided information about *relative availability of evidence,* classified in four levels as "very poor/lacking", "poor", "fair" and "good".

2.5.1 Starting level of evidence and decision flow in data assessment.

The strongest quality of evidence needs to ground on randomized clinical trials performed in subjects undergoing and not undergoing FDG-PET-based diagnosis, and comparing relevant clinical outcomes (i.e., patients' health, survival, quality of life, costs), as the dependent variables, in the two conditions. These studies are currently not available in the FDG-PET literature. In order to allow evidence-based decisions for the use of diagnostic tests, assessment of the quality of accuracy studies of test performance is deemed acceptable (10, 11), provided that such studies have a good starting quality and strong methodology, i.e., they must report validated measures of test performance and perform head to head comparison between the index test and the comparator. Moreover, evidence must exist, linking test performance to patient outcomes (10, 11). Again, in the field of FDG-PET such evidence is not systematically available. The lack of these key elements prevents demonstration of both utility of the exam or lack thereof (Table 3), and thus prevents to derive any decision to support clinical use from evidence.

We have then assessed FDG-PET diagnostic studies as to any other quality item that was available and pertinent to assess the starting quality of evidence and risk of bias, in order to provide anyway information of the current status of the available literature. We thus assessed study design, the presence of quantitative measures of test performance (constituting the critical outcomes in our assessment) of the index test (FDG-PET), the adequacy of gold- or reference-standards, and factors

negatively (section 2.5.2) or positively (section 2.5.3) affecting the starting quality of evidence (Table 3).

Gold- or acceptable reference-standards were pathology, biomarker-based diagnosis, confirmation of diagnosis or decline at clinical follow-up; presence of specified mutations was considered gold-standard for familial AD and Huntington disease; clinical diagnosis was considered gold standard for ALS. In our assessment, we extracted and assessed data also for those papers where the reference-standard used to compute the value of critical outcomes was the mere baseline clinical diagnosis. However, we did not consider these studies as providing evidence of diagnostic utility of FDG-PET.

Table 3. *Quality items for deriving decisions from evidence in FDG-PET accuracy studies, defining the starting quality of evidence for FDG-PET studies. Lacks in these items lead to faulty internal validity, and prevents deriving decisions from evidence relative to the clinical use of FDG-PET. Overcoming many of these limitations is practically feasible based on currently available datasets and methodology; we outlined as research priorities (→RP) the most feasible improvements that are advisable for the next studies on FDG-PET diagnostic accuracy. Flow decisions relative to the evidence assessment aimed to the definition of EANM-EAN recommendations are evidenced in bold.*

STUDY FEATURE	SPECIFIC TARGET FOR DIAGNOSTIC STUDIES	SPECIFIC HURDLES FOR FDG-PET	AVAILABILITY	STRATEGY ADOPTED TO PROCEED
Randomized clinical trial design	Assessment of outcomes of downstream management in patients diagnosed with the index test compared to those diagnosed without it (only traditional examination).	The exceedingly high number of confounders (different alternative diagnostic tools; alternative management and treatment strategies; lack of disease modifying treatment), possible ethical issues and costs are relevant hurdles to such studies. In the lack of disease modifiers, rates of mortality, morbidity, institutionalization, direct and indirect costs, quality of life and caregivers burden may be assessed, but hardly attributable to the use of the exam.	No	Assess diagnostic studies of test performance, with the aim to derive decisions on clinical use from this kind of literature, as deemed admissible for diagnostic tests (10),(11).
Observational study design comparing test performance versus traditional diagnosis (Accuracy studies)	Performance of the index test quantitatively compared head-to-head versus the comparator using the same gold-standard.	No hurdles are envisioned for comparing directly the performance of FDG-PET and traditional diagnosis. The lack of this features denotes faults in the internal validity of available studies. This prevents demonstration of both utility of the index test or lack thereof.	Very rare	**Decisions on clinical use of FDG-PET cannot be based on evidence. We proceed with expert consensus. We anyway assessed literature as to any available quality item** to provide panelists, with the highest amount of evidence obtainable from available studies. (→**RP**: performance of comparator can be assessed based on available datasets) When head-to-head comparison was available, studies were considered as having high starting quality.
Linked evidence	Evidence linking test results to patient outcomes (implies availability of effective treatment).	Similar to the RCT issue (first line of this table), cannot be obtained reliably and with reasonable effort in the absence of disease modifiers. Estimates are costly and unreliable due to high variability of downstream patient management.	No	We postulate that better diagnosis leads to better delivery of care.
Validated measures of test performance	Quantitative assessment of the performance in detecting or excluding a target disease.	No hurdles. Measures typically include sensitivity, specificity, accuracy, ROC, AUC, positive and negative predictive value, positive and negative likelihood ratios.	Not always available. Often faulty due to faulty design (mainly inadequate reference standard)	These measures were considered as critical outcomes in the evidence assessment procedure. However, the lack of appropriate gold standard and of head-to-head comparison of index test and comparator performances limit their validity. **Being the only quantitative evidence available in the field, we considered these as critical outcomes for evidence assessment.** (→RP provision of proper measures of test performance is feasible)
Appropriate gold standard for test performance assessment	Pathology or biomarker-based diagnosis (gold-standard) or clinical follow-up (accepted reference standard) allow a proper	Autopsy is usually very difficult to obtain in dementia studies. Biomarker-based diagnosis is available in few recent studies (→RP: use biomarkers in next studies). However, whilst providing a good proxy for pathology, it is currently limited by partial completion of the	VERY RARE	**Decisions on clinical use of FDG-PET cannot be based on evidence, and need to be defined through expert consensus.** We considered diagnosis at follow-up as a proxy for appropriate gold-standard.

assessment of test performance, being independent on both the index test and comparator.	validation process in the dementia field (22). Although limited, the use of, at least, clinical follow up is highly recommended, in the lack of stronger gold-standard (→RP).

2.5.2 Factors negatively affecting the starting level of evidence

Factors negatively affecting the starting quality of evidence relate to biases preventing to remove the effect of confounders. They are either the same as for intervention studies or not pertinent to FDG-PET studies (table 4). Further specifications are reported for the outcomes below.

Lack of blinding: to the aim of assessing FDG-PET utility, it is required that scan readers be blind to clinical, diagnosis and gold-standard information for the examined patients.

Use of non-validated outcome measures: Many studies reported only patterns of hypometabolism associated to the target disease, resulting from regression analyses or t-test comparison of patients versus controls. These are considered as "typical" for the examined disease, but do not provide any quantification of univocal correspondence with the target diagnosis, nor any measures of test accuracy, and were thus considered as providing no evidence of utility, although data were anyway extracted and presented. Exceptions were PICOs 18 and 20, based on the specific target of the PICO question. When available, measures of clinical outcomes like change in diagnosis, diagnostic confidence, treatment and prediction of survival were included in our assessment as proxies.

Indirectness: sources for indirectness of evidence relating to the ultimate link with patient outcomes and to the lack of direct comparison with the comparator were already accounted for in the assessment of the starting quality of evidence. Among factors negatively affecting the starting quality of evidence in terms of indirectness, we thus considered only the differences between the study population, intervention or outcomes of interest reported in the study compared to those addressed in the pertinent PICO question. These included sample features limitedly representative of the target population due to different severity, age at onset, ethnic group; the use of semiquantitative methods of image analysis as to intervention, as such methods are not yet widespread in clinical routine; or kinds of comparisons that did not match exactly those required in the PICO.

Publication Bias. We did not perform formal analyses, but rather considered the starting quality of evidence to be negatively affected based on a possible publication bias in the cases when more papers providing the same results were from the same research group or dataset.

2.5.3 Factors positively affecting the quality of evidence

Large effect: we considered the effect large for test performance values between 81-100%, medium between 71-80%, and small between 51-70%.

D*ose-response gradient* is not pertinent to FDG-PET, and the possibility that *confounding factors* could hide an important effect was not assessable due to the exceedingly large number of methodological limitations that may concur in blurring the target effect (Table 4).

2.6 Summaries of evidence assessment

In order to facilitate the communication on available evidence with panelists, we produced, besides the tables with all extracted data for each PICO, summary of findings tables reporting the outcomes assessed globally across papers, and abstracts better elucidating all findings (see (23-29). These explained both the quantitative data as based on the methodological assessment described, and the "qualitative" findings, as based on the pattern of hypometabolism correlated to the target disorders for each PICO. Panelists were informed that this final ranking of relative availability of evidence across the 21 PICOs should not have been considered in the same terms as the absolute quality of evidence, as it can be provided based on Cochrane systematic reviews or GRADE assessment. The very frequent lack of the basic methodological requirements of the available studies was clearly reported, as well as the lack of any evidence of utility of studies describing metabolic patterns significantly associated to the target condition but not characterized quantitatively with validated measures of diagnostic performance.

2.7 Delphi procedure

The Delphi procedure (30) is already described elsewhere (7). Briefly, the panel was formed uniquely by clinicians, expert in FDG-PET, nominated within the EANM and EAN. They answered the 21 PICO questions that they defined at the beginning of the project using a web-based platform. Panelists were asked to decide considering both the literature so assessed and their own expertise, and to provide the reasons for their decisions writing them in mandatory windows within the voting system. At each round, they could access statistics on answers from the previous rounds and the anonymized answers and justifications provided by the other panelists. Questions were considered answered, and not re-proposed in further rounds, after a majority of at least 5 vs 2 was achieved.

Table 4. *Quality items for deriving decisions from evidence: factors negatively and positively affecting the starting quality of evidence as from (5, 12).*

QUALITY ITEMS FOR DERIVING DECISIONS FROM EVIDENCE	AIM FOR DIAGNOSTIC STUDIES	PERTINENCE TO FDG-PET	FDG-PET-SPECIFIC
	FACTORS NEGATIVELY AFFECTING THE QUALITY OF EVIDENCE		
LACK OF ALLOCATION CONCEALMENT	Same as for intervention studies	Not applicable at present due to the lack of RCT	
LACK OF BLINDING	Test readers should be blind to clinical and gold/reference standard results	Pertinent	Assessed as "no risk" if blinding is reported, or "unclear risk" if information about blinding is not reported
INCOMPLETE ACCOUNTING OF PATIENTS AND OUTCOMES EVENTS	In diagnostic studies may be considered in relation to the follow-up diagnosis when used as reference standard	Little pertinent: often, the reference standard is follow-up diagnosis, and its availability defines the sample	Not systematically available
SELECTIVE OUTCOME REPORTING	Analogous to intervention studies	Pertinent	To allow metanalyses, values should be reported also for non-significant results. Includes failure in reporting outcomes that should be reported.
EARLY INTERRUPTION FOR BENEFIT	Not applicable	Not pertinent	
USE OF NON VALIDATED OUTCOME MEASURES	e.g., measures other than quantitative assessment of test accuracy	Pertinent	"Typical" patterns of hypometabolism associated to the target condition, but devoid of quantitative assessment of diagnostic performance were not considered as providing evidence of utility. Measures of incremental diagnostic value or treatment change considered as proxies.
CARRYOVER EFFECT	Analogous to intervention studies	Not pertinent	Not applicable
RECRUITMENT BIAS	Analogous to intervention studies	Pertinent	Assessed with reference to consecutive recruitment

INCONSISTENCY	Analogous to intervention studies	Pertinent	Assessed in the global outcome evaluation across studies. We converged here the assessment of imprecision, based on the analysis of confidence intervals.
IMPRECISION	Analogous to intervention studies	Pertinent	Different from intervention studies, we treated the imprecision due to large confidence intervals within the "effect inconsistency" domain. We treated imprecision based on the information provided for scan features.
INDIRECTNESS	a) (Indirectness as to the link with ultimate patient outcomes) b) (Indirectness due to lack of direct comparison between index-test and comparator) c) indirectness due to differences in population, intervention, outcome of interest: analogous to intervention studies	a-b) already accounted for in the starting quality of evidence c) pertinent	c) Examples for FDG-PET can be: patients with atypical presentation compared to healthy controls rather than to patients potentially entering differential diagnosis; patients with subjective cognitive decline (SCD) compared versus healthy controls, rather than SCD progressing to objective cognitive decline versus non-converter, etc. Correlation studies may be considered "indirect as to outcome of interest", however quality is further lowered by circularity in this case; this issue is already accounted for in study design (correlation studies not appropriate) and in the use of non-validated outcome measures.
PUBLICATION BIAS	Same as for intervention studies	Pertinent	may require analyses overcoming the heterogeneity of measures of test performance, not performed.
	FACTORS POSITIVELY AFFECTING THE QUALITY OF EVIDENCE		
LARGE EFFECT	Analogous to intervention studies.	Pertinent	We considered 51-70% = small, 71-80% = medium, 81-100% = large effect
DOSE-RESPONSE GRADIENT	Not pertinent	Not pertinent	Not applicable
CONFOUNDING FACTORS	Analogous to intervention studies	Pertinent	Difficult to assess due to frequent coexistence of many methodological problems

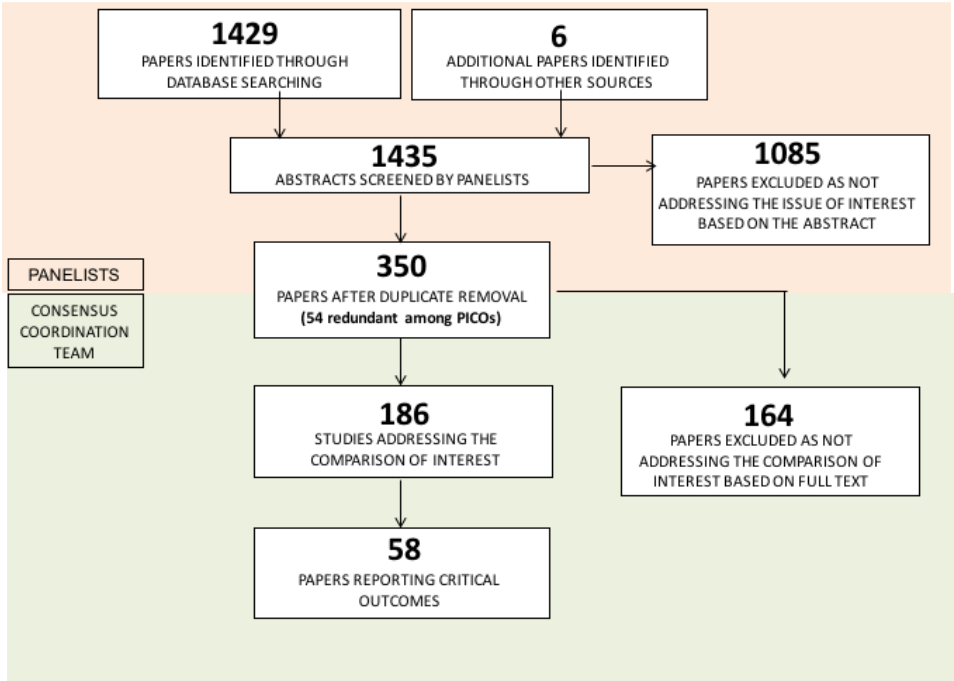
3 Results

More specific information on the results reported in this section is provided elsewhere (23-29).

3.1 Literature selection.

For the 21 searches, a total of 1435 papers was identified and screened for subsequent processing. After excluding papers not addressing the comparison of interest and duplicate papers, a total of 186 papers was assessed in greater detail to evaluate the quality of evidence (Figure 1).

Figure 1. *Literature search performed to assess evidence supporting the use of FDG-PET for the 21 PICO questions detailed in Table 1. The search and a first screening was performed by the group of panelists, clinicians expert of FDG-PET from EANM and EAN. An independent methodology team extracted the data and assessed the evidence. 186 papers addressed the comparisons of interest for the 21 PICOs. Among these, only 58 reported the critical outcomes allowing to assess evidence.*



Of these, most (128) did not provide validated measures of test performance, reporting only the patterns of hypometabolism associated to the disease of interest based usually on correlation analyses or t-test comparison with the specific control groups. While 6 PICOs (2,3,5,12,14,18) lacked entirely any critical outcome, only 31% of the total set of papers including the comparison of interest did report proper quantification in terms of accuracy, AUC, predictive value or likelihood ratios (Table 5). Among these, only 2 studies, both pertaining to PICO 21, performed a head-to-head

comparison of FDG-PET versus clinical comparison. Sixty-two percent of the included studies assessed the performance of the index-test versus an acceptable gold- or reference-standard (green or yellow boxes in Table 6).

3.2 Studies reporting critical outcomes

The 14 PICOs reporting critical outcomes had a minimum of 1 study and a maximum of 13, the median being one paper. The total number of subjects per PICO with critical outcomes was very variable, from 13 for PICO 6 to 1361 for PICO 1. Most studies with critical outcomes reported values of accuracy. Only a minority (N=21/58 papers) reported positive and negative predictive values or likelihood ratios. As well, effect sizes had large variability, ranging from 38% to 100% for sensitivity, 41-100% for specificity, and 58-100% for accuracy (Table 5). Pico-specific values are reported in detail in the specific reviews in this issue (7, 23-29)

Table 5. *Summary of the index test assessments for the PICOs reporting critical outcomes. GR1: target population; GR2: comparison group; AUC: Area Under Curve; PPV: Positive Predictive Value; NPV: Negative predicative value; LH+: positive likelihood ratio; LH-: negative likelihood ratio. For PICO 16, specific additional critical outcomes are accuracy indices in detecting the molecular pathologies underlying the PPA syndrome (e.g., amyloidosis or tauopathies). Ach: Acetylcholinesterase inhibitors. NA= not applicable or not available.*

PICO	TOTAL N OF STUDIES	CRITICAL OUTCOME	N STUDIES	SAMPLE SIZE GR1	SAMPLE SIZE GR2	MIN EFFECT	CI	MAX EFFECT	CI
PICO 1	13	Sensitivity	10	388	545	38.00	8-75%	98.00	87-100%
		Specificity	10	388	545	41.00	26-56%	97.00	82-100%
		Accuracy	11	395	555	58.40	43-74%	100.00	80-100%
		PPV	6	224	317	41.00	22-59%	85.20	75-92%
		NPV	6	224	317	77.00	CI: 64-87	95.00	75-100%
		AUC	7	245	421	66.00	CI NA	96.50	91-100%
		LR+	1	77	50	8.14	4.75–13.96		
		LR-	1	77	50	0.12	0.06–0.23		
PICO 6	2	Sensitivity	2	13	30	100.00	59-100%	100	85-100%
		Specificity	2	13	30	83.00	36-100%	100	59-100
		Accuracy	2	13	30	97.00	82-100%	100	77-100%
PICO 7	4	Change in diagnosis	1	37	0	59.50			
		Change in patient management	1	37	0			increase Ach da 13.8 a 38.3%	
		Sensitivity	1	6	27	83.00	36-100%		
		Specificity	1	6	27	85	52-98%		
		Accuracy	1	6	27	83.00	59-96%		
		AUC	3	79	0	82	NA	91	NA
PICO 8	11	Sensitivity	9	156	360	70	47-87%	92	61-100%
		Specificity	9	156	360	74	57-88	100	73-100%
		Accuracy	10	176	380	72	60-82%	96	92-98%
		AUC	5	117	312	77.1	NA	97	NA
		PPV	1	30	37	86	66-95%		
		NPV	1	30	37	85	69-94%		
		LR+	1	30	37	4.46	2.16–9.20		
PICO 9	5	Sensitivity	4	312	173	80	67-89%	99	96-100%
		Specificity	4	312	173	63	35-85%	98	87-100%
		Accuracy	4	253	135	87	69-96%	89.2	75-96%
		AUC	2	261	107	0.91	0.85-0.97	0.97	NA
		PPV	1	62	45	98	88-100%		
		NPV	1	62	45	74	59-86%		

		LR+	1	62	45	29.88	11.61-40.00	
		LR-	1	62	45	0.25	0.13-0.40	
		Other outcomes (logistic regression results)	1	27	24	beta=1.414		
PICO 10	1	Sensitivity	1	27	98	71.00	50-86	
		Specificity	1	27	98	65	55-75%	
		Accuracy	1	27	98	66	57-75%	
		AUC	1	27	98	68	NA	
PICO 11	1	Sensitivity	1	51	51	100	93-100%	
		Specificity	1	51	51	100	93-100%	
		Accuracy	1	51	51	100	96-100%	
PICO 13	2	Sensitivity	2	36	32	52.9	28-77%	75	49-91%
		Specificity	2	36	32	80	56-94%	100	73-100%
		Accuracy	2	36	32	67.6	50-82%	83.9	66-94%
		AUC	1	19	12	80	NA	
PICO 15	2	Sensitivity	2	39	0	91	59-100%	95	NA
		Specificity	2	39	0	58	29-82%	82	
		Accuracy	2	39	0	73	51-88%	82	
		PPV	1	14	0	68	NA	
		NPV	1	14	0	97	NA	
		LR+	1	14	0	3.9	NA	
		LR-	1	14	0	0.06	NA	
PICO 16	4	Sensitivity	1	15	14	86.2	68-96%	
		Specificity	1	15	14	66.7	9-99%	
		Accuracy	1	15	14	84	67-95%	
		PPV	1	15	14	96.1	80-100%	
		NPV	1	15	14	33.3	4-78%	
		Detect AD pathology	2	53	0			
		Detect non-AD pathology	2	31	0			
PICO 17	2	Sensitivity	2	265	60	94.8	86-98%	95.4	91.4-97.9%
		Specificity	2	265	60	80	56-94%	82.5	67.2-92.7%
		Accuracy	2	265	60	91.8	83-96%	93.2	89.2-96.5%
PICO 19	1	AUC	1	26	17	94	NA	
PICO 20	1	Significant correlation between frontal hypometabolism	1	8	NA			

		and cognitive performance							
PICO	TOTAL N OF STUDIES	CRITICAL OUTCOME	N STUDIES	SAMPLE SIZE GR1	SAMPLE SIZE GR2	VISUAL ASSESSMENT		SEMI-QUANTITATIVE ASSESSMENT	
						MIN EFFECT	MAX EFFECT	MIN EFFECT	MAX EFFECT
PICO 21	9	Incremental value indices	3	156	157				
		Sensitivity	6	479	126	59	89.6	62.3	96
		Specificity	6	479	126	50	96	84	99
		Accuracy	7	459	237	64.8	89.2	70	97.5
		AUC	3	155	142	50	87.8	67	96.7
		PPV	3	294	167	68	87.5	84.2	98
		NPV	3	294	167	72	92.4	71	89
		LR+	4	382	279	1.55	14.8	6.08	36.5
		LR-	4	382	279	0.12	0.45	0.03	0.41

3.3 Relative availability of evidence

Considering overall the quality of methodology (type of gold- or reference-standard and head-to-head comparison between index-test and comparator), the number of total subjects for PICO, effect sizes, and consistency of results across studies (see (23-29)), PICO 21 resulted the only one with relatively good evidence. PICO 21 had a total of 9 studies with validated measures of accuracy as proper critical outcomes, with a total of 586 subjects, and considerably high and consistent values of effect size (Tables 5 and 6; (28)); for this PICO, only 3 studies had inadequate reference standard (only baseline clinical diagnosis)(28). The relatively higher strength of PICO 17 on ALS was due to the fact that baseline clinical diagnosis was considered an appropriate gold standard in this specific case (23).

Lower total number of subjects, strength, and consistency of results (as from Table 5 and (23-29)) were observed progressively for the PICOs receiving lower ranking as from table 6.

Table 6. Availability of evidence for the 21 PICO questions assessed. *Numbers denote the number of papers assessing FDG-PET (index test) and traditional clinical diagnosis (comparator) performance in detecting the target disorder, versus gold- and reference-standard, or versus mere baseline clinical diagnosis. Colors in the comparison columns denote methodological appropriateness; in the Outcome columns colors summarize both methodological appropriateness and strength of results as from Table 5 and (23-29) (dark green=good, light green: fair; yellow=poor; red=very poor/lacking). Evidence availability for PICO 17 was considered good although FDG-PET performance was tested*

against mere baseline clinical diagnosis, since this was considered as the proper reference standard for diagnosing ALS.

PICO	INDEX TEST (FDG-PET)			CRITICAL OUTCOME (+PROXY)	COMPARATOR (CLINICAL/NPS ASSESSMENT)			OUTCOME
	Pathology/ Biomarker	Diagnosis/ conversion at fu	Baseline clinical diagnosis		Pathology/ Biomarker	Diagnosis/ conversion at fu	Baseline Clinical Diagnosis	
21- SCAN ASSESSMENT	2	3	4	9	1	1	-	2
8 – DLB vs AD	2	-	9	11	-	-	-	0
17 – ALS	-	-	2	2	-	-	-	0
1 – MCI due to AD	-	13	-	13	-	-	-	0
9 – AD vs FTLD	1	1	3	5	-	-	-	0
15 – CBS	2	-	-	2	-	-	-	0
6 – ADAD	2	-	-	2	-	-	-	0
7 – Atypical AD	2	1	1	4	-	-	1	1
16 – PPA	3	-	1	4	-	-	-	0
19 – pre-HD	-	5	-	1+4	-	-	-	0
20 – HD	1	-	-	1	-	-	-	0
4 – SCI	-	-	1	0+1	-	-	-	0
10 – DLB vs FTLD	-	-	1	1	-	-	-	0
11 – AD vs VaD	-	-	1	1	-	-	-	0
13 – PSP vs IPD	-	1	1	2	-	-	-	0
18 – ALS	-	-	-	0	-	-	-	0
12 – PD related decline	-	-	-	0	-	-	-	0
5 – At risk for AD	-	-	-	0	-	-	-	0
2 – MCI due to FTLD	-	-	-	0	-	-	-	0
3 - MCI due to DLB	-	-	-	0	-	-	-	0
14 – Dep. Pseudo-dementia	-	-	-	0	-	-	-	0
TOTAL	15 (24%)	24 (38%)	24 (38%)	58+5				

3.4 Research priorities

Based on the considerations done while making decisions on evidence assessment, we could spot, besides the many limitations typical of FDG-PET literature, many aspects that can be importantly improvement in the short term. In particular:

Head-to-head comparison with the comparator. All but two studies reporting critical outcomes lack the direct comparison between the index test and the comparator, required for proper methodology as the starting level of evidence. All of these studies, that do assess the accuracy of FDG-PET-based diagnosis versus an appropriate gold- or reference-standard, can use the same reference to assess the performance of the baseline clinical diagnosis, independently on the FDG-PET results, and thus compare this performance with that of the FDG-PET-based diagnosis, to

provide the measure of the incremental diagnostic value. Such values can already be computed using the very same datasets already used to produce the published results. This information would immediately provide a measure of the *incremental* value of FDG-PET-based diagnosis as compared to traditional clinical and neuropsychological diagnosis, that is currently lacking entirely, and would importantly increase the quality of the available evidence.

Computation of measures of test performance independent on the prevalence of the disease in the population. Most studies provide measures of test performance that are widely accepted, although they are critically dependent on the prevalence of the disease in the examined population (i.e., sensitivity, specificity, accuracy). The same datasets employed to produce the published data allow also the computation of other measures of test performance, i.e. positive and negative predictive values and likelihood ratios, that are independent on disease prevalence, and thus critically more informative to clinicians.

Biomarker-based diagnosis. To date, the possibility to perform pathophysiological diagnoses with biomarkers is increasingly concrete. Changes in diagnosis based on information of brain amyloidosis can amount to about 30% (31), thus this kind of improvement should be highly recommended for next papers on FDG-PET diagnostic performance whenever possible.

Discussion

In this study we have detailed the evidence assessment procedure performed for FDG-PET diagnostic studies to the aim of producing European recommendations (7) for prescribing the exam in the diagnostic work-up for dementing neurodegenerative disorders. From the inclusive literature searches performed for 21 PICO questions, we found a minority of papers reporting validated measures of test performance, and thus eligible for proper evidence assessment according to currently acknowledged methodology (8, 17, 18, 21). Among these, almost all did not comply with basic requirements of internal validity. We outline that while, to date, some parameters cannot be computed within reasonable costs or time limits (e.g., the assessment of FDG-PET-based diagnosis on patient outcome in randomized trials, in the absence of disease modifiers), many other relevant requirements may easily be complied with, based on the information already available in currently used datasets. We have thus outlined research priorities addressing very feasible and significant improvements of evidence quality in the short term. Whilst interim decisions on clinical use of FDG-PET have been taken based on currently available data and panelists' expertise (7), such improvements may allow to derive decisions directly from evidence, consistent with current methodological guidance, in a hopefully near future (5).

On the whole, the work of evidence assessment performed within this EANM-EAN initiative is partly analogous to previous efforts. In a literature assessment based on GRADE and including a large metanalysis (32), high values were obtained for PET imaging in providing early and differential diagnosis of AD, indicating better performance of automated versus visual assessment. However, it is difficult to assess similarities and differences with our assessment procedure due to lack of details about assessment and decisions on how to evaluate evidence quality in the absence of basic methodological requirement, as from (17, 18, 21), in (32). In the Cochrane review examining the evidence of FDG-PET utility in supporting the diagnosis of AD in MCI (33), the assessment was based on QUADAS-2 (9), according to the Cochrane methodology (8, 17). In that case, and different from (32), both the selection of papers and the final results were very similar to those obtained in our work (24), and evidence was considered insufficiently strong to support clinical use. Similarly, our PICO 1, relating to the ability of FDG-PET to support AD diagnosis in MCI, ranked relatively well compared to the availability of evidence of the 21 PICOs (Table 6); however, also in our assessment, the *absolute* quality of evidence was low, as shown by the large range of values per outcome and wide confidence intervals in Table 5, and by the "Low" quality assigned to most outcomes (see the last column of the summary of findings "Table PICO 1" in (24)).

To our knowledge, this work is new in its detailing the specific way each item of evidence quality was used for the specific field of diagnostic studies for neurodegenerative diseases. Previous analogous papers, even those addressing the need to adapt traditional methods of evidence assessment to the less sound field of diagnostic studies, did not treat with this detail each specific item of evidence quality, and never related specifically to FDG-PET nor other biomarker specific to neuroimaging nor to dementia (10, 11, 34, 35). Moreover, our effort is new in detailing the methodological strengths and weaknesses of current FDG-PET diagnostic studies, with the aim to concretely address methodological improvement of next studies.

Short-term feasible improvements

a) Quantitative assessment of test performance. In the clinical field, the reliance on so called "typical patterns of hypometabolism" is widespread. However, the circular overlap of metabolic patterns with clinical syndromes, together with the limited correspondence of syndromes and pathology (36) makes studies lacking validated measures of test performance useless to the aim of deriving decisions relative to clinical use. Moreover, even when providing validated measures of test performance, most studies limit the analysis to accuracy, sensitivity and specificity. The computation of values more indicative of test performance independent on the prevalence of the disorder in the population, such as PPV-NPV or likelihood ratios, is rare, but very valuable to support its clinical utility, and can be provided based on currently available datasets.

b) Quantitative assessment of incremental diagnostic value. The lack of any quantification of performance for the comparator, i.e. traditional clinical-neuropsychological diagnosis, prevents the quantification of the *incremental* value of FDG-PET over traditional work-up. This may have relatively little importance to clinicians, searching for independent confirmation and using the exam as an add-on to clinical assessment. However, a formal assessment may also outline a detrimental or null value, and is necessary to clinical as well as to policy decisions on diagnostic procedures and health refunders. Currently, changes in diagnosis, diagnostic confidence and treatment are frequently provided in clinical studies, included FDG-PET (37, 38); although accepted due to current constraints, they do not clarify whether the exam does allow formulating a more exact diagnosis.

Medium-term methodological improvements

a) Gold standard. On the other hand, other methodological requirements, also necessary to the proper outline of evidence-based guidelines (18, 39), may not be equally achievable in the short term. These include for example the difficulty in obtaining pathological specimens as the gold

standard. The use of biomarker-based diagnoses to date may be relatively weak due to their still incomplete validation (22), however this would represent a very valuable improvement, in the lack of pathology confirmation.

b) Evidence from randomized clinical trials (RCT). The most proper evidence allowing to draw decisions requires that randomized clinical trials assess clinically relevant outcomes in patients diagnosed with or without FDG-PET. Besides the high costs and the likely lack of potential sponsors for FDG-PET, sharing a similar destiny as the "orphan drugs", proper completion of RCTs requires rather univocal treatment courses downstream to diagnosis, and possibly the availability of disease modifiers. Methodological adaptations have been proposed that may allow deriving decisions from data lacking RCT-derived evidence for diagnostic tests in evolving fields like those of dementia (10, 11, 34, 35); however, the basic methodological requirements outlined in this paper as short term priorities can be complied with and should no longer be disregarded, in order for these adaptations to be properly adopted.

The many methodological problems in the available literature may not be allowing a clinical utility of FDG-PET to emerge, nor can such literature demonstrate any lack of utility of the exam. These negative findings should be read as a lack of evidence, both in support or against clinical use of FDG-PET, and proper evidence still needs to be collected in future studies. On the other hand, the criticism may be moved that our PICOs addressed too specific questions, that cannot be properly answered by FDG-PET, i.e., the attempt to identify underlying pathophysiology of a variety of neurodegenerative disorders. FDG-PET is indeed acknowledged as a biomarker of downstream neurodegeneration and progression of neurodegenerative diseases. Despite this, current diagnostic criteria do recommend its use as a supportive feature for different kinds of neurodegenerative disorders (16, 40-45) mainly because in several conditions pathophysiological biomarkers are to date under development, at a very different stage. Whereas alpha-synuclein (CSF and tissue biopsy) and Tau biomarkers (CSF and PET) are a step forward, others are much less advanced, such as those for TDP-43. At present, thus, the topographic patterns of hypometabolism in those conditions lacking the possibility to be confirmed with pathological biomarkers keep being an important guide to clinicians in approaching the right diagnosis. This explains both our definition of the 20 clinical PICOs, and the reason why the seven experts from the two Societies considered FDG-PET to provide clinical utility in most symptomatic neurodegenerative conditions, consistent with the perceived utility of clinicians (7).

We propose this work as a systematic input to help improve current accuracy studies in the field, and to better focus future efforts in achieving evidence-based guidelines for the use of FDG-PET.

Acknowledgements

The procedure for assessing scientific evidence and defining consensual recommendations was funded by the European Association of Nuclear Medicine (EANM) and by the European Academy of Neurology (EAN). We thank the Guidelines working group of EAN, particularly Simona Arcuti and Maurizio Leone, for methodological advice.

References

1. Frisoni GB, Perani D, Bastianello S, Bernardi G, Porteri C, Boccardi M, et al. Biomarkers for the diagnosis of Alzheimer's disease in clinical practice: an Italian intersocietal roadmap. Neurobiology of aging. 2017;52:119-31.
2. Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, et al. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. The Lancet Neurology. 2014;13(6):614-29.
3. Jack CR, Jr., Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. The Lancet Neurology. 2013;12(2):207-16.
4. Boccardi M, Gallo V, Yasui Y, Vineis P, Padovani A, Mosimann U, et al. The biomarker-based diagnosis of Alzheimer's disease. 2-lessons from oncology. Neurobiology of aging. 2017;52:141-52.
5. 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. European journal of neurology. 2013;20(3):410-9.
6. Leone MA, Keindl M, Schapira AH, Deuschl G, Federico A. Practical recommendations for the process of proposing, planning and writing a neurological management guideline by EAN task forces. European journal of neurology. 2015;22(12):1505-10.
7. Nobili F, Arbizu J, Bouwman FH, Drzezga A, Agosta F, Nestor P, et al. EANM-EAN recommendations for the use of brain 18F-Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) in neurodegenerative cognitive impairment and dementia: Delphi consensus. European journal of neurology. 2018;being submitted.
8. Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. Chapter 9: Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 100: The Cochrane Collaboration; 2009.
9. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36.
10. Hsu J, Brozek JL, Terracciano L, Kreis J, Compalati E, Stein AT, et al. Application of GRADE: making evidence-based recommendations about diagnostic tests in clinical practice guidelines. Implementation science : IS. 2011;6:62.
11. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ (Clinical research ed). 2008;336(7653):1106-10.
12. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). Journal of clinical epidemiology. 2011;64(4):407-15.
13. Jack CR, Jr., Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, et al. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. The Lancet Neurology. 2010;9(1):119-28.
14. Dubois B, Feldman HH, Jacova C, Cummings JL, Dekosky ST, Barberger-Gateau P, et al. Revising the definition of Alzheimer's disease: a new lexicon. The Lancet Neurology. 2010;9(11):1118-27.
15. Dubois B, Hampel H, Feldman HH, Scheltens P, Aisen P, Andrieu S, et al. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. Alzheimer's & dementia : the journal of the Alzheimer's Association. 2016;12(3):292-323.

16. McKeith IG, Boeve BF, Dickson DW, Halliday G, Taylor JP, Weintraub D, et al. Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium. Neurology. 2017;89(1):88-100.
17. Bossuyt PM, Leeflang MM. Developing Criteria for Including Studies. In: Collaboratioh TC, editor. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 042008.
18. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. Journal of clinical epidemiology. 2011;64(4):383-94.
19. Morbelli S, Garibotto V, Van De Giessen E, Arbizu J, Chetelat G, Drezgza A, et al. A Cochrane review on brain [(1)(8)F]FDG PET in dementia: limitations and future perspectives. European journal of nuclear medicine and molecular imaging. 2015;42(10):1487-91.
20. Matias-Guiu JA, Cabrera-Martin MN, Garcia-Ramos R, Moreno-Ramos T, Valles-Salgado M, Carreras JL, et al. Evaluation of the new consensus criteria for the diagnosis of primary progressive aphasia using fluorodeoxyglucose positron emission tomography. Dementia and geriatric cognitive disorders. 2014;38(3-4):147-52.
21. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. Journal of clinical epidemiology. 2011;64(4):401-6.
22. Frisoni GB, Boccardi M, Barkhof F, Blennow K, Cappa S, Chiotis K, et al. Strategic roadmap for an early diagnosis of Alzheimer's disease based on biomarkers. The Lancet Neurology. 2017;16(8):661-76.
23. Agosta F, Altomare D, Festari C, Orini S, Gandolfo F, Boccardi M, et al. Clinical utility of FDG-PET in amyotrophic lateral sclerosis and Huntington disease. European journal of nuclear medicine and molecular imaging. 2018(In this issue).
24. Arbizu J, Festari C, Altomare D, Walker Z, Bouwman FH, Rivolta J, et al. Clinical utility of FDG-PET for the differential diagnosis in MCI. European journal of nuclear medicine and molecular imaging. 2018(In this issue).
25. Bouwman FH, Orini S, Gandolfo F, Altomare D, Festari C, Agosta F, et al. FDG-PET in the differential diagnosis between different forms of PPA. European journal of nuclear medicine and molecular imaging. 2018(In this issue).
26. Drzezga A, Altomare D, Festari C, Arbizu J, Orini S, Herholz, et al. Clinical utility of FDG-PET in the evaluation of conditions at risk for AD. European journal of nuclear medicine and molecular imaging. 2018(In this issue).
27. Nestor P, Altomare D, Festari C, Drzezga A, Rivolta J, Walker Z, et al. Clinical utility of FDG-PET for the differential diagnosis among the main forms of dementia. European journal of nuclear medicine and molecular imaging. 2018(In this issue).
28. Nobili F, Festari C, Altomare D, Agosta F, Orini S, Van Laere, et al. Automated assessment of FDG-PET for the differential diagnosis in patients with neurodegenerative disorders. European journal of nuclear medicine and molecular imaging. 2018(In this issue).
29. Walker Z, Gandolfo F, Orini S, Garibotto V, Agosta F, Arbizu J, et al. Clinical utility of FDG-PET in Parkinson’s disease and atypical Parkinsonisms associated to dementia European journal of nuclear medicine and molecular imaging. 2018(In this issue).
30. Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CF, Askham J, et al. Consensus development methods, and their use in clinical guideline development. Health technology assessment (Winchester, England). 1998;2(3):i-iv, 1-88.
31. Barthel H, Sabri O. Clinical Use and Utility of Amyloid Imaging. Journal of nuclear medicine : official publication, Society of Nuclear Medicine. 2017;58(11):1711-7.

32. Perani D, Schillaci O, Padovani A, Nobili FM, Iaccarino L, Della Rosa PA, et al. A survey of FDG- and amyloid-PET imaging in dementia and GRADE analysis. BioMed research international. 2014;2014:785039.
33. Smailagic N, Vacante M, Hyde C, Martin S, Ukoumunne O, Sachpekidis C. (1)(8)F-FDG PET for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). The Cochrane database of systematic reviews. 2015;1:Cd010632.
34. Gopalakrishna G, Mustafa RA, Davenport C, Scholten RJ, Hyde C, Brozek J, et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. Journal of clinical epidemiology. 2014;67(7):760-8.
35. Trenti T, Schunemann HJ, Plebani M. Developing GRADE outcome-based recommendations about diagnostic tests: a key role in laboratory medicine policies. Clinical chemistry and laboratory medicine. 2016;54(4):535-43.
36. Beach TG, Monsell SE, Phillips LE, Kukull W. Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005-2010. Journal of neuropathology and experimental neurology. 2012;71(4):266-73.
37. Laforce R, Jr., Buteau JP, Paquet N, Verret L, Houde M, Bouchard RW. The value of PET in mild cognitive impairment, typical and atypical/unclear dementias: A retrospective memory clinic study. American journal of Alzheimer's disease and other dementias. 2010;25(4):324-32.
38. Laforce R, Jr., Tosun D, Ghosh P, Lehmann M, Madison CM, Weiner MW, et al. Parallel ICA of FDG-PET and PiB-PET in three conditions with underlying Alzheimer's pathology. NeuroImage Clinical. 2014;4:508-16.
39. GRADE handbook for grading quality of evidence and strength of recommendations2013.
40. Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & dementia : the journal of the Alzheimer's Association. 2011;7(3):270-9.
41. Gorno-Tempini ML, Hillis AE, Weintraub S, Kertesz A, Mendez M, Cappa SF, et al. Classification of primary progressive aphasia and its variants. Neurology. 2011;76(11):1006-14.
42. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Jr., Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & dementia : the journal of the Alzheimer's Association. 2011;7(3):263-9.
43. Rascovsky K, Hodges JR, Knopman D, Mendez MF, Kramer JH, Neuhaus J, et al. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain : a journal of neurology. 2011;134(Pt 9):2456-77.
44. Hoglinger GU, Respondek G, Stamelou M, Kurz C, Josephs KA, Lang AE, et al. Clinical diagnosis of progressive supranuclear palsy: The movement disorder society criteria. Movement disorders : official journal of the Movement Disorder Society. 2017;32(6):853-64.
45. Strong MJ, Abrahams S, Goldstein LH, Woolley S, McLaughlin P, Snowden J, et al. Amyotrophic lateral sclerosis - frontotemporal spectrum disorder (ALS-FTSD): Revised diagnostic criteria. Amyotrophic lateral sclerosis & frontotemporal degeneration. 2017;18(3-4):153-74.