Diagnostic Accuracy of PET for Recurrent Glioma Diagnosis: A Meta-Analysis

T. Nihashi, I.J. Dahabreh, and T. Terasawa

ABSTRACT

BACKGROUND AND PURPOSE: Studies have assessed PET by using various tracers to diagnose disease recurrence in patients with previously treated glioma; however, the accuracy of these methods, particularly compared with alternative imaging modalities, remains unclear. We conducted a meta-analysis to quantitatively synthesize the diagnostic accuracy of PET and compare it with alternative imaging modalities.

MATERIALS AND METHODS: We searched PubMed and Scopus (until June 2011), bibliographies, and review articles. Two reviewers extracted study characteristics, validity items, and quantitative data on diagnostic accuracy. We performed meta-analysis when ≥5 studies were available.

RESULTS: Twenty-six studies were eligible. Studies were heterogeneous in treatment strategies and diagnostic criteria of PET; recurrence was typically suspected by CT or MR imaging. The diagnostic accuracies of 18F-FDG (n = 16) and 11C-MET PET (n = 7) were heterogeneous across studies. 18F-FDG PET had a summary sensitivity of 0.77 (95% CI, 0.66–0.85) and specificity of 0.78 (95% CI, 0.54–0.91) for any glioma histology; 11C-methionine PET had a summary sensitivity of 0.70 (95% CI, 0.50–0.84) and specificity of 0.93 (95% CI, 0.44–1.00) for high-grade glioma. These estimates were stable in subgroup and sensitivity analyses. Data were limited on 18F-FET (n = 4), 18F-FLT (n = 2), and 18F-boronophenylalanine (n = 1). Few studies performed direct comparisons between different PET tracers or between PET and other imaging modalities.

CONCLUSIONS: 18F-FDG and 11C-MET PET appear to have moderately good accuracy as add-on tests for diagnosing recurrent glioma suspected by CT or MR imaging. Studies comparing alternative tracers or PET versus other imaging modalities are scarce. Prospective studies performing head-to-head comparisons between alternative imaging modalities are needed.

ABBREVIATIONS: CI — confidence interval; 11C-MET — 11-carbon methionine; 18F-FET — 18-fluorine fluoroethyltyrosine; 18F-FLT — 18-fluorine fluorothymidine; ROC — receiver operating characteristic; 201TI — thallium 201

Gliomas are the most commonly diagnosed primary brain tumors in the United States.1 Despite recent advances in temozolomide-based multimodality therapy, high-grade gliomas remain incurable diseases with a median survival of <5 years for glioblastoma and <5 years for anaplastic gliomas.2–4 In contrast, low-grade gliomas, which include astrocytoma, oligodendroglioma, and oligoastrocytoma, are indolent malignant tumors, typically surgically treated,3,5 with a median survival of >5 years.

Treatment-induced necrosis is a common treatment-related morbidity in the management of gliomas, which typically occurs 3–12 months posttreatment.6 On conventional MR imaging with gadolinium enhancement, treatment-induced necrosis typically presents as an increase in the size of contrast-enhancing lesions, which mimics tumor progression or recurrence after remission.

Received May 1, 2012; accepted after revision July 29.

From the Department of Radiology (T.N.), Nagoya University Graduate School of Medicine, Nagoya, Japan; Center for Clinical Evidence Synthesis (I.J.D., T.T.), Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts; Center for Evidence-based Medicine (I.J.D.), Brown University, Providence, Rhode Island; Department of Internal Medicine (T.T.), Fujita Health University Nanakiri Sanatorium, Fujita Health University School of Medicine, Tsu, Mie, Japan; and Clinical Research Center (T.T.), National Hospital Organization Nagoya Medical Center, Nagoya, Japan.

Takashi Nihashi and Teruhiko Terasawa contributed equally to the study and manuscript.

Please address correspondence to Teruhiko Terasawa, MD, PhD, Fujita Health University Nanakiri Sanatorium, 424-1 Odoricho, Tsu, Mie, 514-1295 Japan; e-mail: terasawa@fujita-hu.ac.jp

Evidence-Based Medicine Series 1

http://dx.doi.org/10.3174/ajnr.A3324

Copyright 2012 by American Society of Neuroradiology.
Differentiating the 2 conditions is challenging, and reliable noninvasive neuroimaging modalities are needed to better guide the management of patients with suspected recurrence.

PET is a promising molecular neuroimaging technique that provides metabolic tumor information complementing the CT and MR imaging examinations. Several studies have evaluated PET by using various tracers (eg, 18F-FDG, 11C-MET, 18F-FET, or 18F-FLT) as a test for aiding the differential diagnosis of suspected glioma recurrence. 18F-FDG is the most widely used tracer; its uptake correlates with the amount of glucose consumption and the local metabolic rate within the glioma lesion.9 Uptake of 18F-FDG in high-grade glioma is typically similar to or less than that in normal gray matter; uptake in low-grade glioma is similar to that in white matter.8,10 Due to the low contrast between tumor and healthy brain tissue with 18F-FDG, however, more specific tracers have been developed. Amino acid tracers such as 11C-MET and 18F-FET offer higher contrast than 18F-FDG based on the increased intracellular amino acid use and extracellular matrix production of tumor cells.8,9 Further, uptake of 18F-FLT correlates well with thymidine kinase-1 activity, a cytosolic enzyme with high concentration in proliferating cells but low in resting cells.8,10 Because cell proliferation rates are higher in malignant glioma cells compared with scar tissue, 18F-FLT can also differentiate tumor recurrence from treatment-induced necrosis.

Studies evaluating novel PET tracers have small sample sizes and use heterogeneous designs, making interpretation of published data difficult. Furthermore, the comparative effectiveness of alternative imaging modalities such as advanced MR imaging techniques (eg, perfusion MR imaging or MR spectroscopy) is currently uncertain. We performed a systematic review to provide a comprehensive summary and quantitative synthesis of information on the diagnostic accuracy of PET by using various tracers to diagnose disease recurrence in patients with previously treated glioma. We also aimed to compare PET with other imaging modalities for differentiating recurrent or progressive glioma from treatment-induced necrosis, when used as add-on tests to conventional MR imaging.

MATERIALS AND METHODS
Search Strategy, Study Eligibility, and Data Abstraction
We searched the Medline and Scopus databases (from inception through June 30, 2011) with no language restriction. The complete search strategies are presented in the On-line Appendix. To complement our data base searches, we examined the reference lists of eligible studies and relevant review articles.

Two reviewers (T.N., T.T.) independently screened abstracts and further examined full-text articles of potentially eligible citations. Studies that assessed PET by using any tracer for differentiating disease recurrence from treatment-induced necrosis in patients with suspected glioma recurrence after any form of treatment were eligible. We included both prospective and retrospective studies, and we considered pathologic confirmation with or without clinical follow-up as the reference standard. We included only English language publications that evaluated at least 10 patients; smaller studies do not provide meaningful estimates of accuracy. We excluded studies that did not provide adequate information to allow the calculation of sensitivity and specificity. We also excluded editorials, comments, letters to the editor, and review articles.

One of 2 reviewers (T.N., T.T.) extracted descriptive data from each eligible study, which were verified by a second reviewer. We extracted the following information from eligible studies: first author, year of publication, journal, patient demographic and clinical characteristics, therapeutic interventions, technical specifications of PET, and interpretation of PET results. Two reviewers (I.J.D., T.T.) independently extracted quantitative data regarding imaging results and final diagnoses. Discrepancies were resolved by consensus. When studies performed a direct comparison between different imaging modalities (eg, 18F-FDG PET versus thallium 201 [201TI] SPECT), we extracted data on accuracy for all imaging tests investigated.

We took particular care to identify publications with at least partially overlapping populations by comparing authors, centers, recruitment periods, patient demographic characteristics, and glioma histologies. We included all relevant publications in qualitative synthesis but only included studies with nonoverlapping patient populations in meta-analyses, to avoid double counting of evidence. Specifically, when multiple publications with potentially overlapping patient populations were available, we only included the publication with the largest sample size in the meta-analysis.

Validity Assessment
To assess the validity and reporting quality of studies, we evaluated 14 items that were considered relevant to the review topic on the basis of the Quality Assessment of Diagnostic Accuracy Studies instrument.11,12 The complete operational definition of each item is available from the authors on request. For comparative studies of diagnostic tests, we extracted the proportion of study participants receiving each comparator test. We operationally defined an “optimal direct comparison” as the performance of both tests at the same time point in at least 90% of eligible patients. This cutoff was chosen to limit the potential for patient selection and disease progression bias. Two reviewers (T.N., T.T.) independently assessed study quality, and discrepancies were resolved by consensus.

Data Synthesis
For each study, we constructed a 2 × 2 contingency table consisting of true-positive, false-positive, false-negative, and true-negative results. Patients were categorized according to whether they were test positive or negative (on the basis of imaging) and whether they had relapsed glioma by the reference standard. We extracted results of visual and quantitative assessments separately. When a study reported test results at multiple time points during clinical follow-up, we only recorded the results of the test performed closest to the completion of treatment (ie, the first instance of PET performance after recurrence was suspected). Also, in studies in which histologic results were negative and clinical follow-up results were also reported, we planned to only consider the clinical status as the reference standard because it is more important from the patient’s perspective. However, no cases of discrepant results between pathologic and clinical reference standards were reported in the studies we reviewed.
We recorded the counts of true-positive, false-positive, false-negative, and true-negative results based on the cutoff values specified by each study (when reported). When studies did not specify cutoff values but did report numeric data of quantitative assessment for each enrolled patient, we used the following methods to construct a $2 \times 2$ of test results: for $^{18}$F-FDG, we determined the optimal cutoff threshold for defining positive and negative scans by ROC analysis; for $^{11}$C-MET results, we used a cutoff value of 1.5 of the tumor-to-normal reference ratio or similar indices to define positive (>1.5) and negative ($\leq$1.5) results as the main analysis as recommended by experts and the optimal cutoff threshold determined by ROC analysis as a sensitivity analysis.

We calculated sensitivity and specificity for each study with their corresponding 95% CIs. We obtained summary estimates of sensitivity and specificity with their corresponding 95% CIs by using bivariate random effects meta-analysis with the exact binomial likelihood, when at least 5 studies were available (because of model complexity at least 5 studies are required for estimation). Summary positive and negative likelihood ratios were calculated from the summary sensitivity and specificity estimates. We assessed between-study heterogeneity visually, by plotting sensitivity and specificity separately in forest plots, and also in the ROC space. We constructed summary ROC curves and confidence regions for summary sensitivity and specificity when appropriate. Analyses were conducted by using STATA 11.1/SE (StataCorp, College Station, Texas) and Meta-Analyst, Version 3.0 (Tufts Medical Center, Boston, Massachusetts). All tests were 2-sided, and statistical significance was defined as a $P$ value < .05.

**RESULTS**

**Study Selection and Characteristics**

Our PubMed and Scopus searches identified 2808 and 3835 citations, respectively, of which 48 were considered potentially eligible and were retrieved in full text for further assessment. We identified an additional potentially eligible article by perusing the reference lists of relevant review articles. After full text review, 23 publications were excluded and 26 studies were considered eligible for this review (Fig 1). A complete list of excluded studies along with reasons for exclusion is provided in the Appendix.

To explore heterogeneity, we performed a subgroup analysis limited to high-grade gliomas only. To further explore whether study-level characteristics could explain between-study heterogeneity, we performed univariat (single predictor) meta-regression analyses by using the bivariate model (2 outcomes, sensitivity and specificity, modeled jointly). We assessed the following, a priori selected, covariates: year of publication, study design (prospective versus retrospective), study size, relapse rate, proportion of use of temozolomide, and type of reference standard (pathology only versus pathology and clinical follow-up).

Analyses were conducted by using STATA 11.1/SE (StataCorp, College Station, Texas) and Meta-Analyst, Version 3.0 (Tufts Medical Center, Boston, Massachusetts). All tests were 2-sided, and statistical significance was defined as a $P$ value < .05.
were \(^{18}\text{F-FDG}\) (16 studies) and \(^{11}\text{C-MET}\) (7 studies). Four studies compared different PET tracers: Two compared \(^{18}\text{F-FDG}\) with \(^{11}\text{C-MET},^2,^3\) and 1 compared \(^{18}\text{F-FLT}\) with \(^{18}\text{F-FET}.^4\) Six studies reported comparisons between \(^{18}\text{F-FDG}\) PET and other imaging modalities: 2 with \(^{201}\text{Tl-SPECT},^7,^8\) 2 with perfusion MR imaging,\(^9,^3\) 1 with iodine 123-alfa-methyl-tyrosine \(^{201}\text{TI-SPECT},^1\) and 1 with MR spectroscopy.\(^3\) Four studies compared \(^{11}\text{C-MET}\) PET with other imaging modalities: 2 with perfusion MR imaging,\(^9,^3\) 1 with \(^{201}\text{Tl-SPECT},^2\) and 1 with MR spectroscopy.\(^3,^3\)

Twenty of 26 studies (77%) included both high- and low-grade gliomas,\(^6,^12,^34-^40\) and 6 exclusively included high-grade gliomas.\(^34-^39\) Although few studies reported specific clinical contexts of PET assessment (ie, whether included patients were investigated for potential recurrence after primary therapy or after salvage therapy), patients typically underwent PET > 6 months after completion of therapy. Twenty-one of 26 studies (77%) had a retrospective design. Typically, the reference standard comprised biopsy or clinical follow-up; 4 of 26 studies (15%) used biopsy as the sole reference standard to assess recurrence. Mean or median follow-up ranged between 7 and 34 months.

Studies adopted various treatment strategies (On-line Table 2). Most used multimodality therapies typically comprising surgery and some form of radiation therapy with or without chemotherapy. Only 4 recent studies explicitly reported that patients with high-grade glioma were treated with temozolomide.\(^33,^38-^40\)

Regarding imaging techniques and technologies, included studies generally followed guidelines by the European Association of Nuclear Medicine (On-line Table 3).\(^42,^43\) All studies used stand-alone dedicated PET scanners except for 1 study, in which some patients underwent combined PET/CT instead of stand-alone PET.

Studies used variable diagnostic criteria both for visual and quantitative assessments (On-line Table 4). The 2 indexes most commonly used in quantitative assessments were maximum standardized uptake values within the region of interest of the suspected lesions and the ratio of uptake in the suspected lesion to that in a reference area. No study explicitly reported how the region of interest was specified. Studies typically reported pairs of sensitivity and specificity on the basis of the optimal cutoff values estimated by ROC analysis. Only 1 study reported inter-rater agreement when multiple interpreters were involved in the interpretation of PET results.\(^25\)

**Assessment of Validity**

No study adequately reported all 14 items relevant to study validity that we assessed (On-line Table 5). Reporting was particularly poor regarding the following items: blinding of interpreters of the index and reference standard tests, whether the decision to perform biopsy was based on PET results, and whether additional treatments were applied during clinical follow-up. Most studies had a retrospective design and did not clearly report whether consecutive patients were included. Six of 26 studies (23%) adopted a reference standard comprising biopsy only without clinical follow-up.

**Sensitivity, Specificity, Likelihood Ratios, and Summary ROC Curves**

Studies reported heterogeneous sensitivities and specificities for \(^{18}\text{F-FDG PET}\) (On-line Fig 1). When considering both low- and high-grade gliomas,\(^16-^19,^21-^23,^27,^32,^33,^40,^41\) sensitivity ranged between 0.23 and 0.95 and specificity ranged between 0.17 and 1.0. For high-grade gliomas,\(^16-^19,^21-^23,^27,^34,^35,^39,^41\) sensitivity ranged between 0.18 and 1.00 and specificity ranged 0.25 and 1.0. Studies including both low- and high-grade gliomas had a summary sensitivity of 0.77 (95% CI, 0.66 – 0.85) and a summary specificity of 0.78 (95% CI, 0.54 – 0.91), corresponding to a positive likelihood ratio of 3.4 (95% CI, 1.6 – 7.5) and a negative likelihood ratio of 0.30 (95% CI, 0.21– 0.43) (Fig 2). In analyses limited to high-grade glioma, summary sensitivity was 0.79 (95% CI, 0.67 – 0.88) and summary specificity was 0.70 (95% CI, 0.50 – 0.84), corresponding to a positive likelihood ratio of 2.6 (95% CI, 1.5 – 4.4) and a negative likelihood ratio of 0.30 (95% CI, 0.20– 0.46). The summary ROC curves and confidence regions for summary sensitivity and specificity were similar for studies of both all glioma histologies and the subgroup of high-grade gliomas with comparable Q* statistics of 0.77 and 0.75, respectively. These estimates were similar in subgroup analyses in which only studies adopting visual assessment were considered and in sensitivity analyses (On-line Fig 2).

Similarly, studies reported widely ranging sensitivity and specificity estimates for \(^{11}\text{C-MET PET}\). When we considered both low- and high-grade gliomas,\(^20,^25,^29,^30\) sensitivity ranged between 0.55 and 0.80 and specificity ranged between 0.70 and 1.0 (On-line Fig 1). For high-grade gliomas,\(^20,^25,^30,^36,^39\) sensitivity ranged between 0.44 and 0.93 and specificity ranged between 0.50 and 1.0. For high-grade gliomas, the summary sensitivity was 0.70 (95% CI, 0.50 – 0.84) and summary specificity was 0.93 (95% CI, 0.44– 1.00), corresponding to a positive likelihood ratio of 10.31 (95% CI, 0.76–139.39) and a negative likelihood ratio of 0.32 (95% CI, 0.18–0.57), and the Q* statistic was 0.79 (Fig 2). These estimates were similar in subgroup analyses in which only studies using quantitative assessment were considered and in sensitivity analyses (On-line Fig 2).

Studies evaluating \(^{18}\text{F-FET},^2,^4,^25,^26,^28,^31\) and \(^{18}\text{F-FLT},^31,^40\) reported consistently high sensitivity, particularly for high-grade glioma (On-line Fig 1). We did not perform meta-analysis for \(^{18}\text{F-FET},^2,^4,^25,^26,^28,^31\) \(^{18}\text{F-FLT},^31,^40\) or \(^{18}\text{F-boronophenylalanine},^37\) because there were few studies investigating each of these tracers.

**Meta-Regression Analyses**

We performed meta-regression analyses only for studies evaluating \(^{18}\text{F-FDG PET}\) (both for all glioma histologies and for high-grade glioma alone) because this was the only tracer evaluated in > 10 studies. In meta-regression analyses, year of publication, study design, sample size, relapse rate, proportion of use of temozolomide, or the type of reference standard did not affect test performance statistically significantly (all \(P\) values > .05).
Comparisons among Different PET Tracers
Six studies \(^{20,25,31,36,39,40}\) reported on 9 comparisons among pairs of different PET tracers. In all except 1 study, both tests were performed in at least 90% of study participants. Generally, test comparisons showed trade-offs between sensitivity and specificity (consistent with diagnostic threshold effects), suggesting that different tracers may have broadly similar diagnostic accuracy (Online Fig 3).

Comparisons between PET and Other Imaging Modalities
Six studies \(^{17,21,27,32,33,39}\) reported on 10 comparisons between 18F-FDG PET and other imaging tests. Five comparisons were between 18F-FDG PET and advanced MR imaging techniques; only 1 of these comparisons involved >90% of study participants. No study explicitly stated how patients were selected for additional diagnostic testing, suggesting that selection bias may have affected results. ROC plots did not show any clear pattern (Online Fig 3).

Four studies \(^{20,30,38,39}\) reported on 8 comparisons between 18C-MET PET and other imaging tests. Among 4 comparisons between 11C-MET PET with advanced MR imaging techniques, only 1 involved >90% of study participants. Studies again did not report why patients were excluded from additional testing. No consistent pattern regarding comparative diagnostic accuracy was evident from these studies (Online Fig 3).

DISCUSSION
This systematic review suggests that PET by using 18F-FDG or 11C-MET has moderately good overall accuracy for diagnosing disease recurrence, independent of histologic grade, among patients with glioma for whom recurrence was suspected by conventional anatomic imaging tests such as CT or MR imaging. These results are mainly based on visual assessment for 18F-FDG, and quantitative assessment for 11C-MET; however, various diagnostic criteria and thresholds were adopted across studies. Evidence on other tracers is sparse. Furthermore, evidence is limited regarding comparisons among different PET tracers, as well as for the comparison of PET with non-PET imaging modalities.

18F-FET and 18F-FLT are relatively new PET tracers and have been in clinical use only for the past decade. Their diagnostic accuracy has been assessed only in a limited number of referral centers. Although promising pilot data have been reported, further validation is needed. The studies we reviewed typically focused on individual imaging modalities and did not allow a complete comparative evaluation of the available imaging technologies. Furthermore, most studies retrospectively and jointly assessed low- and high-grade gliomas treated with heterogeneous treatment strategies and adopted heterogeneous methodologies for confirming disease relapse with diverse follow-up protocols. Our findings expand the findings of previous narrative reviews \(^{9,44}\) by providing a quantitative overview of the diagnostic accuracy of PET for differentiating between disease recurrence and treatment-induced changes. Additionally, our work provides a comprehensive review of studies directly comparing different PET tracers and those comparing PET with alternative imaging tests in this clinical setting.

Several limitations of the available evidence are worth noting. Most studies had limited internal and external validity; therefore, reported accuracy estimates may not be replicable or relevant to other clinical settings. Also, few studies used treatment strategies that would be consistent with the current standards of care (ie, surgery alone for low-grade glioma and temozolomide-based multimodality therapy for high-grade glioma). Thus, our results may be less applicable to current clinical practice. In addition, our analyses on comparative evidence are based on a limited number of studies. Furthermore, studies comparing PET with MR imag-
ing modalities are based mostly on selected patients; therefore, our results should be interpreted with caution. Finally, no studies of $^{18}$F-choline or $^{3,4}$-dihydroxy-$^{18}$F-fluoro-L-phenylalanine PET fulfilled our inclusion criteria, and we did not consider non-comparative studies of non-PET imaging modalities.

Pseudoprogression is a clinically benign phenomenon characterized by the appearance and subsequent stabilization (or spontaneous regression) of enhancing lesions on routine MR imaging, within 2 months after completion of concurrent treatment with temozolomide and radiation. It is unclear whether our results are directly applicable to the use of PET for differentiating between pseudoprogression and true tumor progression; few patients in our review had been treated with temozolomide-based therapies, and most patients had been evaluated with PET at later time than when pseudoprogression is typically suspected.

Future studies of PET for the evaluation of glioma relapse should use prospective designs, focus on clinically relevant patient populations treated with standardized protocols, and avoid potential biases in evaluating test accuracy. An important research priority is the assessment of test performance for distinguishing pseudoprogression from true tumor progression in the context of temozolomide-based treatment for high-grade glioma. Given the current emphasis on the comparative effectiveness of health care interventions, research efforts should focus on the relative benefits and risks of competing imaging modalities in real-world clinical settings. Data on head-to-head comparisons among individual imaging modalities (ie, comparisons among different PET tracers [eg, $^{18}$F-FDG PET versus $^{11}$C-MET PET] and comparisons of PET versus novel MR imaging techniques [eg, MR spectroscopy, diffusion-weighted imaging, and perfusion-weighted imaging]) and more complex testing strategies (eg, PET alone versus PET plus another non-PET technique) are particularly needed.

CONCLUSIONS

Both $^{18}$F-FDG and $^{11}$C-MET PET have moderately good overall accuracy for detecting recurrent glioma in patients with suspected recurrence following active treatment. Data on other PET tracers, though seemingly promising, are scarce. Comparative evidence is generally limited and whether a specific PET tracer outperforms others or whether PET is superior to alternative imaging modalities remains unclear. Prospective comparative studies are needed to elucidate the optimal imaging strategy for evaluating patients with suspected recurrent glioma.

ACKNOWLEDGMENTS

We thank Drs Masazumi Fujii and Shigenori Takebayashi (Department of Neurosurgery, Nagoya University Graduate School of Medicine) and Dr Yoshiro Ando (Department of Radiology, Nagoya University Graduate School of Medicine) for helpful comments on a previous version of the manuscript.


REFERENCES