How should we analyse FDG PET studies for monitoring tumour response?

Adriaan A. Lammertsma[1], Corneline J. Hoekstra[2], Giuseppe Giaccone[3], Otto S. Hoekstra[1]

[1] Department of Nuclear Medicine and PET Research, VU University Medical Centre, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands
[2] Department of Nuclear Medicine, Jeroen Bosch Hospital 's-Hertogenbosch, The Netherlands
[3] Department of Medical Oncology, VU University Medical Centre, Amsterdam, The Netherlands

Abstract. FDG PET is a promising technique for monitoring tumour response early during anticancer therapy. Progress, however, has been limited owing to the multitude of methods currently in use. Here, the most promising techniques for multi-centre trials are discussed briefly, with emphasis on the need for standardisation. In addition, an approach is presented for response monitoring studies using newly developed drugs. This approach makes use of a large database of response monitoring studies, which defines the relationship between simplified clinical methods and full quantitative analysis for classic cytotoxic drugs. For a new drug, first a pilot study is performed to assess whether it affects this relationship. Based on this pilot, it is decided whether or not a simplified clinical method can be used in further studies.

Keywords: Multi-centre trials – Standardisation – Response monitoring – FDG PET

Introduction

The ultimate goal of anticancer therapy is to kill all tumour cells whilst completely sparing all normal tissues. In practice, this is never possible. Especially for systemic therapies, such as chemotherapy, the dose delivered will be a trade-off between maximum tumour kill and minimum side-effects. This dose is based on results from large clinical trials, but there is never a guarantee that it will be effective in an individual patient. As the number of anticancer drugs increases, it is becoming more and more important to develop sensitive tools to distinguish responders from non-responders early during or, preferably, prior to therapy.

The early detection of non-response is important for individual patients, especially if alternative drugs or therapies are available. It allows switching to such an alternative therapy without the patient being exposed to the potentially severe toxic effects of the first ineffective drug. Moreover, if this alternative therapy is effective, there is an additional benefit in that delay in starting effective therapy is minimal. Even if there is no alternative (potentially) curative therapy, the patient will be spared from exposure to a toxic, but useless, therapy and one could focus on proper palliative treatment.

Early discontinuation of ineffective treatment is also beneficial for society. Some of the newer anticancer drugs are extremely expensive. From a financial point of view it becomes impossible to prescribe these drugs to all patients with a given tumour. There is increasing debate about who should qualify for a certain drug, and some patients have already turned to a court of law to demand treatment. Clearly, an objective method that enables the early selection of responders and non-responders would have the additional benefit of maximising resources for those patients who will benefit from a treatment.

Within the context of drug development, a method to distinguish responders from non-responders early during (or even prior to) treatment would be beneficial. Development of drugs that are ineffective could be discontinued before starting expensive clinical trials. On the other hand, such a method might identify subgroups of patients who will benefit from the new drugs. Especially when such a subgroup is small, this would have gone unnoticed in a larger trial. In other words, the method might provide a means to compose more homogeneous patient groups.

Prediction of response

Ideally, efficacy of a drug should be determined prior to any treatment. A prerequisite for efficacy is bioavailability,

Adriaan A. Lammertsma (✉)
Department of Nuclear Medicine and PET Research,
VU University Medical Centre,
P.O. Box 7057,
1007 MB Amsterdam, The Netherlands
e-mail: aa.lammertsma@vumc.nl
Tel.: +31-20-4444214, Fax: +31-20-4443090

i.e. the concentration in the tumour should be high enough for a kill effect, whilst it should remain below that level in normal tissues. Bioavailability can be investigated using positron emission tomography (PET). By labelling a drug with a positron emitter, it is possible to measure uptake of that drug at tracer (i.e. non-therapeutic and non-toxic) levels in vivo, both in the tumour and in normal tissues [1]. This is perhaps less interesting for new drugs as development of a labelling strategy might be time consuming for many drugs (complex molecules). Nevertheless, this approach could be extremely useful for an existing (routinely used) drug, especially if it is expensive and the fraction of non-responders is high. In addition, it provides a means for studying the effects of interventions aimed at improving targeting (e.g. manipulating the blood–brain barrier in the case of brain tumours). Clearly, many issues still need to be addressed, such as the complicating issue of potential uptake of labelled metabolites. In addition, it needs to be demonstrated that kinetics are similar at tracer and therapeutic doses.

Early response monitoring using FDG

As mentioned above, methods to predict response prior to treatment are not routinely available yet. Based on the general availability of $[^{18}F]$-2-fluoro-2-deoxy-D-glucose (FDG) for diagnostic purposes, to date most (early) response monitoring studies also have focussed on this metabolic tracer.

The diagnostic success of FDG PET is based on the increased glycolytic rate seen in most tumours [2]. As a result, identification of positive lymph nodes and/or metastases is simple, as they result in areas with increased FDG uptake (hot spots) compared with the low uptake in surrounding normal tissues. For this identification, it is not important whether FDG uptake is proportional to the degree of malignancy. In the case of response monitoring studies, however, it is assumed that the level of uptake is somehow proportional to tumour "activity", i.e. it is assumed that a reduction in uptake following therapy reflects lower tumour "activity" and vice versa. Amongst other things, this means that inherently it is assumed that the lumped constant [3], which accounts for the differences between natural glucose and FDG, remains constant during therapy.

The use of FDG PET for staging only requires visual assessment of whole-body scans (hot spot detection). In some cases this might even be useful or sufficient for assessing response to chemotherapy, as is typically the case for malignant lymphoma after completed therapy. For example, if, on follow-up whole-body scans, new sites with increased uptake are detected, progressive disease is likely. Even for the tumour itself, visual assessment might be all that is needed. If a tumour disappears during treatment (increased uptake in tumour at baseline is absent on follow-up scan), it can be assumed that there has been a good response. A more sophisticated analysis is unlikely to lead to another conclusion. In many cases, however, there will

be some reduction in uptake and some form of quantification will be required to distinguish response from non-response. In addition, a quantitative method has the important advantage that, in theory, it allows for the definition of objective cut-off values for response or, more likely, the definition of response probabilities.

Methods for FDG response monitoring studies

If decisions have to be based on FDG PET studies, be it go/no go for drug development or the treatment strategy for an individual patient, the first principle should be that the measure derived from repeat FDG scans accurately reflects glucose metabolism (as mentioned above, this implies that the lumped constant remains constant, an assumption that applies to all methods of analysis). It can be argued that for initial FDG studies with a new drug, the most accurate technique should always be used, independent of complexity. For larger multi-centre trials and for clinical studies, however, there is a need for simple scanning protocols that are not too time and labour consuming. Unfortunately, this has led to the use of a large number of methods with different levels of complexity [4]. Note that even for a given method, various alternatives have been described (e.g. different normalisation factors for the SUV). As a result it is almost impossible to compare results from different institutes and, more importantly, published cut-off values are method and often institute specific, especially since they are also affected by acquisition protocol, reconstruction algorithm and region of i0nterest definition.

At the VU University Medical Centre (VUmc), a number of response monitoring studies have been performed in lung, breast and oesophagus tumours [5–7]. One purpose of these studies was to compare the large variety of published methods with the view to identify potential candidates for future studies. In all three studies (i.e. tumours), the same four methods, each with a different level of complexity, were identified. Below is a brief description of these four methods, focussing on potential advantages and shortcomings. A full description of these and other methods can be found elsewhere [4].

Standardised uptake value

The most commonly used method is a very simple semi-quantitative method: standardised uptake value (SUV). This method is attractive for clinical studies as it requires only a single (static) scan, usually acquired 60 min after intravenous injection of FDG, and no blood sampling. SUV is then calculated as tumour uptake, normalised to injected dose and a factor that takes into account the total distribution space of FDG. Several factors have been proposed: body weight [8], body surface area [9] and lean body mass [10]. The latter two factors are thought to be more accurate, as they better account for the very low FDG uptake in fatty tissues. To account for possible changes in plasma glucose levels [11], which may occur during

therapy, a correction for plasma glucose has also been proposed. Note that, based on these correction factors, a total of six different definitions are possible. A disadvantage of SUV is its dependency on the time interval between injection and scanning. Therefore, this interval should be kept constant at all times. A further disadvantage is that injected dose (dose calibrator) needs to be measured accurately (cross-calibrated against PET scanner). The main disadvantage, however, is the inherent assumption that plasma clearance is always the same. If plasma clearance of FDG changes as a result of therapy (i.e. due to changing uptake in other tissues), the relationship between uptake at a certain time and injected dose will also change. This cannot be accounted for in the SUV calculation and, consequently, comparison of pre- and post-therapy scans might be misleading.

Non-linear regression

The most accurate, but also most complex, method is non-linear regression (NLR), which takes into account the full two-tissue compartment model describing the kinetics of FDG in tissue [12]. It requires dynamic scanning (multiple frames) from the time of injection. In addition, throughout the scan, continuous or rapid arterial blood sampling is required, unless the heart is in the field of view of the scanner. In the latter case the input function can be derived from the (dynamic) scan itself [13, 14]. Advantages of this method are accurate determination of glucose (actually FDG, but this applies to all methods) metabolism itself, even in the presence of dephosphorylation, the possibility of separately quantifying glucose transporter and hexokinase activity, and, last but not least, the absence of dependency on scanning time or altered plasma clearance. Apart from the complex acquisition protocol, the main limitation of NLR is its sensitivity to noise. Consequently, it is not possible to perform calculation at the voxel level and, therefore, no parametric images of glucose metabolism can be generated. The complex calculations are not really a disadvantage. The software needs to be installed once and calculations can be performed in an automatic way. Only quality control of noisy data (low uptake) might be an issue.

Patlak analysis

The so-called Patlak or Patlak/Gjedde analysis [15] is a linearisation of the compartmental equation for irreversible tracers. This method also requires dynamic scanning, but fast frames over the initial phase are not required. Only a few frames, starting 10 min after injection, are needed. Arterial sampling is still required, but again, there is less demand for rapid sampling. Calculations are faster than for NLR. In addition, they are less sensitive to noise. As a result, it is possible to perform calculations at the voxel level, thereby generating functional images of glucose metabolism [16]. This can be a major advantage for tumours located close to vascular structures, as the contrast between tumour and blood will be much higher than in SUV images. Like NLR, Patlak analysis is independent of changes in plasma clearance of FDG. The method is not valid if there is a significant degree of dephosphorylation, but this is very uncommon in tumours. A disadvantage compared with NLR is the inability to quantify glucose transporter and hexokinase activity.

Simplified kinetic method

A potentially attractive method is the simplified kinetic method (SKM) [17]. It requires only a static scan and a few (venous) blood samples during the scan. These samples are used to scale a population-derived average plasma curve. The main advantage of SKM is its ability to estimate glucose metabolism without the need for a dynamic scan and with a very reduced blood sampling protocol. The advantage over SUV is that it does, at least to some extent, take into account changes in plasma clearance. Disadvantages of SKM are that this correction for differences in plasma clearance is only a first-order correction (the peak is assumed to be constant) and that the method has not been used extensively, i.e. it has not been validated in a large patient population.

Recommendations for standardisation

Multi-centre studies require that methodology be standardised. Within that context it is clear that the multitude of methods that are in use for quantifying FDG uptake has, for many years, hampered widespread use of FDG PET for monitoring response to anticancer therapy. This was first realised by the EORTC PET study group, who took the initiative for standardisation by publishing recommendations for the measurement of response using FDG and PET [18]. These recommendations covered several aspects: patient preparation, scanning protocols, methods of analysis and definition of metabolic response criteria.

The complete EORTC guidelines are beyond the scope of the present review. With respect to quantification, the EORTC defined a minimum standard with the view that all clinical PET groups should be able to adhere to this standard. This minimum standard was defined as SUV corrected for body surface area (SUV_{BSA}), but not for plasma glucose. A correction for plasma glucose was thought not to be necessary, as patients should be fasting for at least 6 h. In addition, such a correction would require an accurate technique rather than the simple glucose assays used by many clinical groups.

The EORTC recommendations also specified that at least one centre should perform a formal comparison of the above-mentioned semi-quantitative SUV results with those obtained from the more quantitative Patlak analysis. The purpose of such a comparison was to verify that, for the combination of tumour and drug under investigation, SUV provided results that were indicative of glucose metabolism.
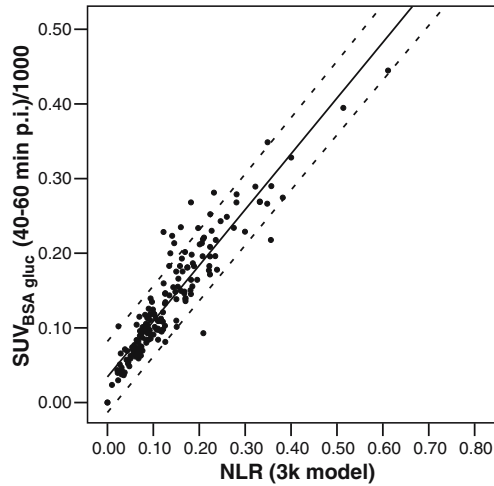
Fig. 1. Regression of $SUV_{BSA,glu}$ (corrected for plasma glucose) versus NLR for a database of 170 scans

More recently, the NCI has drafted new guidelines [19]. Although there are some minor differences and additions compared with the EORTC guidelines, the philosophy of a routine SUV approach, together with an initial comparison with the Patlak method for validation of SUV, remains unchanged.

VUmc approach

In a series of response monitoring studies on lung, breast and oesophageal cancer [5–7], various methods for quantification of FDG uptake were compared. It was shown that Patlak and NLR produced nearly identical results. In addition, the relationship between SUV and Patlak and between SKM and Patlak was constant, suggesting that both simplified methods could be used for monitoring response to anticancer therapy.

By pooling the data from these studies, a database with 170 scans was created, allowing the investigation of correlations between the various methods. As an example, in Fig. 1 the correlation between SUV_{BSA} corrected for plasma glucose and NLR is shown. Intraclass correlation coefficients with corresponding confidence intervals for the agreement between Patlak, SKM and SUV_{BSA}, respectively, versus NLR are shown in Table 1. Although

Table 1. Intraclass correlation coefficients (ICC) with corresponding confidence intervals (CI) for agreement between the method indicated and NLR

Method	ICC	CI
Patlak	0.98	0.97–0.98
SKM	0.94	0.92–0.95
SUV_{BSA}	0.88	0.84–0.91
$SUV_{BSA,glu}$[a]	0.91	0.88–0.93

[a]Corrected for plasma glucose

Patlak provides better correlation with NLR, it remains to be seen whether the lower correlations observed for the other methods have any clinical significance, i.e. the corresponding higher uncertainties need to be compared with the minimum change that still represents a relevant response.

It is clear from these data that even SUV_{BSA} appears to be a good method for monitoring anticancer therapy, at least for the tumours and drugs investigated in the above-mentioned studies [5–7]. For the largest study (lung tumours), there is further evidence that all four methods have similar accuracy for discriminating responders from non-responders when compared with clinical outcome [20]. From the consistency of the data in the database it can be concluded that no substantial effects of classic cytotoxic drugs on the kinetics of FDG have been identified, at least not in settings in which therapy is administered for short periods, ample time is allowed between courses to allow for recovery, and PET scanning is performed after such a recovery period (i.e. just prior to the next course). At present, new drugs are being introduced which are administered continuously. Under these conditions it cannot be assumed that there will be no effects on the kinetics of FDG at the time of repeat scans during therapy.

For response monitoring studies with new drugs at the VUmc, the following procedure is followed. First, a pilot in thoracic tumours is performed with full quantification, i.e. dynamic scanning, image-derived input function and analysis of data using NLR, SKM and SUV_{BSA} with correction for plasma glucose (plasma glucose correction provides slightly better accuracy at minimal cost). Next, the correlations of SKM and SUV_{BSA} with NLR are compared
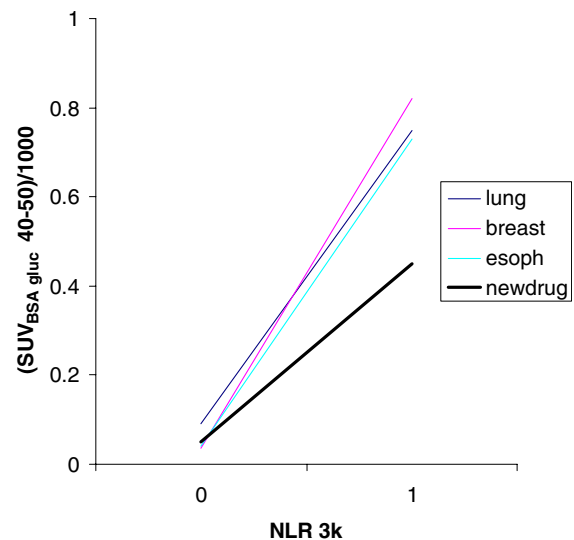

Fig. 2. Regression of $SUV_{BSA,glu}$ (corrected for plasma glucose) versus NLR for a database of 170 scans from lung, breast and oesophageal tumours. Separate regression lines for the three tumours are shown, indicating little variation with tumour type (the overall regression line is shown in Fig. 1). In addition, the regression line is shown for a response monitoring pilot study during treatment with a new drug

with those for the database. If the regression lines do not differ significantly, only the SUV approach (i.e. static scan without blood sampling) is used for the main study. If, on the other hand, there is a significant deviation from the database regression line, then a dynamic scanning protocol will be maintained for the entire study.

The above-mentioned approach is important, especially for newer experimental therapies, where a metabolic effect on other tissues cannot be excluded. In Fig. 2 a recent example is shown in which a SUV-alone approach could give misleading results. Rather than showing a regression line for the entire database, separate lines are shown for lung, breast and oesophagus tumours. It can be seen that there is little difference between these lines. Also shown is the regression line for the correlation between NLR and SUV_{BSA} (corrected for plasma glucose) in lung tumours after treatment with a new experimental drug (for reasons of confidentiality, neither the name of the drug nor the individual data points can be given at present). Of course, before treatment, the relationship corresponded with that of the database. Clearly, after therapy, SUVs were significantly lower than would have been expected on the basis of NLR. In other words, by looking at SUV data alone, in this case one would have substantially overestimated the efficacy of the drug.

Region of interest definition

It is of interest to note that often less attention is paid to the method of defining regions of interest (ROI). Indeed, in published reports on response monitoring, there is as much variety in the methods of ROI definition as there is in data acquisition and quantification. The EORTC, for example, only recommended that the same ROI volumes should be sampled on subsequent scans and that they should be positioned as closely to the original tumour volume as possible. In two recent studies [21, 22] the impact of ROI definition on FDG quantification was investigated using both phantom studies and clinical data. It was shown that consistency in defining ROIs was extremely important for reliable quantitative results. For response studies, the actual type of ROI (maximum, mean, manual) was less important (only a fixed size ROI gave poor results), provided that the same type was used on pre- and post-therapy scans. Clearly, however, an automatic (threshold) technique is the best guarantee that consistent ROIs are defined on repeat scans. For absolute quantification, the type of ROI was very important and an (automatic) threshold technique (e.g. all pixels within a 50% threshold between maximum pixel counts and background) performed best.

Concluding remarks

Evidence is emerging that FDG PET may provide an objective method to discriminate responders from non-responders with obvious applications both for developing new anticancer drugs and for optimising therapy in individual patients. Its real impact, however, needs to be determined in large multi-centre clinical trials. For those studies, standardisation of methods is essential. The optimal method should be a trade-off between accuracy and clinical applicability.

Pooled data presented in the present review demonstrate that for most tumours and drugs, simplified methods for data acquisition and analysis (i.e. SUV-based methods) are sufficient. Nevertheless, care should be taken in applying these simplified methods for the evaluation of new drugs. To avoid misinterpretation of results, it is essential that first the simplified methods are validated against full kinetic analyses.

To date, response monitoring using FDG has focussed on changes in FDG uptake early during therapy. It remains to be seen whether the measurement of changes is sufficient for a full assessment of response. A recent study [20] has demonstrated that residual uptake after one course of chemotherapy might be as important as the change from baseline. It is likely that future response criteria will depend both on change in FDG uptake and on residual uptake early during treatment, as the latter also indirectly accounts for baseline FDG uptake, which in several tumours has been shown to carry prognostic information itself. It should be noted that, for multi-centre studies, standardising the method of ROI definition (and scanner performance) is even more important for the measurement of residual FDG uptake than it is for measuring changes in FDG uptake.

This review only deals with methodology related to the use of FDG PET for response monitoring. As a tracer of metabolism, FDG is not a specific tumour marker. Other tracers (e.g. FLT), probing other tumour characteristics, are in constant development. At present, quantification of clinical data obtained with these tracers can only be achieved by full kinetic analysis. The nuclear medicine community cannot afford to use simplified methods again before they are validated.

References

1. Propper DJ, de Bono J, Saleem A, Ellard S, Flanagan E, Paul J, et al. Use of positron emission tomography in pharmacokinetic studies to investigate therapeutic advantage in a phase I study of 120-hour intravenous infusion XR5000. J Clin Oncol 2003;21:203–210

2. Warburg O. On the origin of cancer cells. Science 1956;123:306–314

3. Sokoloff L, Reivich M, Kennedy C, Des Rosiers MH, Patlak CS, Pettigrew KD, et al. The [^{14}C]deoxyglucose method for the measurement of local cerebral glucose utilization: theory, procedure, and normal values in the conscious and anesthetized albino rat. J Neurochem 1977;28:897–916

4. Hoekstra CJ, Paglianiti I, Hoekstra OS, Smit EF, Postmus PE, Teule GJJ, et al. Monitoring response to therapy in cancer using [^{18}F]-2-fluoro-2-deoxy-D-glucose and positron emission tomography: an overview of different analytical methods. Eur J Nucl Med 2000;27:731–743

5. Hoekstra CJ, Hoekstra OS, Stroobants SG, Vansteenkiste J, Nuyts J, Smit EF, et al. Methods to monitor response to chemotherapy in non-small cell lung cancer with ^{18}F-FDG PET. J Nucl Med 2002;43:1304–1309

6. Krak NC, van der Hoeven JJ, Hoekstra OS, Twisk JWR, van der Wall E, Lammertsma AA. Measuring [^{18}F]FDG uptake in breast cancer during chemotherapy: comparison of analytical methods. Eur J Nucl Med Mol Imaging 2003;30:674–681

7. Kroep JR, Van Groeningen CJ, Cuesta MA, Craanen ME, Hoekstra OS, Comans EFI, et al. Positron emission tomography using 2-deoxy-2-[^{18}F]-fluoro-D-glucose for response monitoring in locally advanced gastroesophageal cancer; a comparison of different analytical methods. Mol Imaging Biol 2003;5: 337–346

8. Wahl RL, Cody RL, Hutchins GD, Mudgett EE. Primary and metastatic breast carcinoma: initial clinical evaluation with PET with the radiolabeled glucose analogue 2-[F-18]-fluoro-2-deoxy-D-glucose. Radiology 1991;179:765–770

9. Kim CK, Gupta NC, Chandramouli B, Alavi A. Standardized uptake values of FDG: body surface area correction is preferable to body weight correction. J Nucl Med 1994;35:164–167

10. Zasadny KR, Wahl RL. Standardized uptake values of normal tissues at PET with 2-[fluorine-18]-fluoro-2-deoxy-D-glucose: variations with body weight and a method for correction. Radiology 1993;189:847–850

11. Lindholm P, Minn H, Leskinen-Kallio S, Bergman J, Ruotsalainen U, Joensuu H. Influence of the blood glucose concentration on FDG uptake in cancer—a PET study. J Nucl Med 1993;34:1–6

12. Phelps ME, Huang SC, Hoffman EJ, Selin C, Sokoloff L, Kuhl DE. Tomographic measurement of local cerebral glucose metabolic rate in humans with (F-18)2-fluoro-2-deoxy-D-glucose: validation of method. Ann Neurol 1979;6:371–388

13. Van der Weerdt AP, Klein LJ, Boellaard R, Visser CA, Visser FC, Lammertsma AA. Image-derived input functions for determination of MRGlu in cardiac ^{18}F-FDG PET scans. J Nucl Med 2001;42:1622–1629

14. Hoekstra CJ, Hoekstra OS, Lammertsma AA. On the use of image-derived input functions in oncological fluorine-18 fluorodeoxyglucose positron emission tomography studies. Eur J Nucl Med 1999;26:1489–1492

15. Patlak CS, Blasberg RG, Fenstermacher JD. Graphical evaluation of blood-to-brain transfer constants from multiple-time uptake data. J Cereb Blood Flow Metab 1983;3:1–7

16. Messa C, Choi Y, Hoh CK, Jacobs EL, Glaspy JA, Rege S, et al. Quantification of glucose utilization in liver metastases: parametric imaging of FDG uptake with PET. J Comput Assist Tomogr 1992;16:684–689

17. Hunter GJ, Hamberg LM, Alpert NM, Choi NC, Fischman AJ. Simplified measurement of deoxyglucose utilization rate. J Nucl Med 1996;37:950–955

18. Young H, Baum R, Cremerius U, Herholz K, Hoekstra O, Lammertsma AA, et al. Measurement of clinical and subclinical tumour response using [^{18}F]-fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. Eur J Cancer 1999;35:1773–1782

19. Shankar LK, Hoffman JM, Bacharach S, Graham M, Karp J, Lammertsma AA, et al. Consensus recommendations on the use of positron emission tomography (PET) and [^{18}F]-2-fluoro-2-deoxy-d-glucose (FDG) as an indicator of therapeutic response in patients involved in national cancer institute (NCI) clinical trials. J Nucl Med 2006; in press

20. Hoekstra CJ, Stroobants SG, Smit EF, Vansteenkiste J, van Tinteren H, Postmus PE, et al. Prognostic relevance of response evaluation using [^{18}F]-2-fluoro-2-deoxy-D-glucose positron emission tomography in patients with locally advanced non-small-cell lung cancer. J Clin Oncol 2005;23:8362–8370

21. Boellaard R, Krak NC, Hoekstra OS, Lammertsma AA. Effects of noise, image resolution and ROI definition on the accuracy of standard uptake values: a simulation study. J Nucl Med 2004;45:1519–1527

22. Krak NC, Boellaard R, Hoekstra OS, Twisk JWR, Hoekstra CJ, Lammertsma AA. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. Eur J Nucl Med Mol Imaging 2005;32:294–301