Correlation between Ki-67 immunohistochemistry and 18F-Fluorothymidine uptake in patients with cancer: A systematic review and meta-analysis

A. Chalkidou a,*, D.B. Landau b, E.W. Odell c, V.R. Cornelius d, M.J. O’Doherty e, P.K. Marsden e

a Comprehensive Cancer Imaging Centre, Imaging Sciences and Biomedical Engineering, Kings College London, 4th Floor, Lambeth Wing, St. Thomas Hospital, SE1 7EH London, UK
b Department of Clinical Oncology, Guy’s and St. Thomas’ NHS Foundation Trust, Lower Ground Floor, Lambeth Wing, St. Thomas Hospital, SE1 7EH London, UK
c Department of Clinical and Diagnostic Sciences, Kings College London Dental Institute, Floor 28, Tower Wing Guy’s Hospital, SE1 9RT London, UK
d Department of Primary Care and Public Health Sciences, Kings College London, 7th Floor, Capital House, 42 Weston Street, SE1 3QD London, UK
e Imaging Sciences and Biomedical Engineering, Kings College London and Guy’s and St. Thomas’ NHS Foundation Trust, Lower Ground Floor, Lambeth Wing, St. Thomas Hospital, SE1 7EH London, UK

Available online 31 May 2012

Abstract  Background: Positron emission tomography (PET) imaging using the radiotracer 18F-Fluorothymidine (FLT) has been proposed as an imaging biomarker of tumour proliferation. If FLT-PET can be established as such it will provide a non-invasive, quantitative measurement of tumour proliferation across the entire tumour. Results from validation studies have so far been conflicting with some studies confirming a good correlation between FLT uptake and Ki-67 score and others presenting negative results.

Methods: Firstly we performed a systematic review of published studies between 1998 and 2011 that explored the correlation between FLT uptake and Ki-67 score and examined possible variations in the methods used. Studies were eligible if they: (a) included patients with cancer, (b) investigated the correlation between Ki-67 measured by immunohistochemistry and FLT uptake measured with PET scanning, and (c) were published as a full paper in a peer-reviewed scientific journal.

Secondly a meta-analysis of the correlation coefficient values reported from each study was performed. Correlation coefficient ($r$) values were extracted from each study and 95% confidence intervals (CIs) were calculated after applying Fisher’s $z$ transformation. For subgroup analysis, studies were classified by the index used to characterise Ki-67 expression (average or maximum expression), the nature of the sample (whole specimen or biopsy) and the cancer type.

* Corresponding author. Tel.: +44 7586819132/02071888377.
E-mail address: anastasia.chalkidou@kcl.ac.uk (A. Chalkidou).
Findings: Twenty-seven studies were identified as eligible for the meta-analysis. In the studies we examined there were variations in aspects of the methods and reporting. The meta-analysis showed that given an appropriate study design the FLT/Ki-67 correlation is significant and independent of cancer type. Specifically subgroup analysis showed that FLT/Ki-67 correlation was high in studies measuring the Ki-67 average expression regardless of use of surgery or biopsy samples \( (r = 0.70, 95\% \text{ CI} = 0.43–0.86, p < 0.001) \). Of the studies that measured Ki-67 maximum expression, only those that used the whole surgical specimen provided a significant \( r \) value \( (r = 0.72, 95\% \text{ CI} = 0.54–0.84, p < 0.001) \). Studies that used biopsy samples for Ki-67 maximum measurements did not produce a significant \( r \) value \( (r = 0.04, 95\% \text{ CI} = -0.18–0.26, p = 0.71) \). In terms of the cancer type subgroup analysis there is sufficient data to support a strong FLT/Ki-67 correlation for brain, lung and breast cancer. No publication bias was detected.

Interpretation: This systematic review and meta-analysis highlights the importance of the methods used in validation studies comparing FLT-PET imaging with the biomarker Ki-67. The correlation is significant and independent of cancer type provided a study design that uses Ki-67 average measurements, regardless of nature of sample, or whole surgical samples when measuring Ki-67 maximum expression. Sufficient data to support a strong correlation for brain, lung and breast cancer exist. However, larger, prospective studies with improved study design are warranted to validate these findings for the rest of the cancer types.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Understanding the molecular biology of cancer as well as identifying cancer therapeutic targets and monitoring response to treatment has conventionally been based upon assays of tissue or blood samples in vitro. Recently imaging biomarkers have been explored as surrogates for these measurements. There are two main advantages associated with imaging biomarkers: the acquisition of comprehensive spatial information across the entire tumour and body, and the ability to perform non-invasive longitudinal assessment. Molecular imaging techniques in particular have shown promising results. However, in order to serve as biomarkers they need to fulfil certain criteria including that they accurately reflect the underlying biological process. This usually requires validation against a reference standard, frequently represented by a tissue biomarker.

The proliferation rate of cancer cells has been proven to have prognostic and predictive value in a number of cancers.12 The positron emission tomography (PET) radiotracer 18F-Fluorothymidine (FLT) has been proposed as an imaging biomarker of proliferation based on its relationship with Thymidine Kinase 1 (TK-1) in the thymidine salvage pathway.3 During S-phase FLT is phosphorylated by TK-1 and trapped inside the cell, but it is not incorporated into DNA, thus it represents an indirect measurement of proliferation.3 A recent search in January 2012 in clinicaltrials.gov using the key words FLT, PET, cancer identified a total of 79 ongoing and recently closed studies evaluating the role of FLT as an imaging proliferation biomarker in cancer.

Preclinical studies have shown strong spatial correlation between FLT uptake and Ki-67 immunohistochemical marker expression (the gold standard for assessing tumour proliferation status in clinical practice).4 For clinical studies however, results have so far been conflicting with some studies confirming a good correlation between FLT and Ki-675–18 and others presenting negative results.19–28 Various biological explanations for the lack of FLT/Ki-67 correlation have been proposed by authors: loss of cell cycle-specific regulation of TK124; cell ATP levels24; FLT representing only the salvage pathway of thymidine metabolism19; diversity of tumour entities in a specific cohort21; difference in the phosphorylation rate between FLT and thymidine.27

In addition to biological causes for the lack of correlation, however, the complex methodology involved in addressing the relationship of a tissue versus an imaging biomarker may also be an important factor and this is frequently ignored. Few authors have suggested that limitations in their study design such as the use of small biopsy samples or estimation of the expression in the areas of highest proliferation activity (Ki-67max) could have an influence on their results.23,24,26 Although Ki-67max has been shown to be clinically relevant in terms of prognosis in a number of tumour sites, when using biopsy samples the accuracy of this measurement will be subject to sampling errors and reduced reproducibility as it does not take into account the degree of intratumour heterogeneity expression for this marker.29 Another approach is to provide an average measurement of Ki-67 expression (Ki-67mean) in an attempt to overcome regional tissue heterogeneity.

PET is a molecular imaging technique that can provide various quantitative measurements of the underlying tumour biology depending on the radiotracer used. The radiotracer is injected through a vein, accumulates in the tumour and the radioactive emissions are detected.
by the PET camera that produces 3D images. The standardised uptake value (SUV) is a simplified measure of tracer uptake, and is the most widely used method for the quantitative analysis of oncological PET studies. Previous studies have highlighted the importance of the many factors that can affect SUV quantification measurements.30

We have therefore performed a systematic review of published studies that explore the correlation between FLT uptake and Ki-67 expression with two objectives: (a) to investigate the variations in methods used in these studies and (b) to perform a meta-analysis of the reported correlation values examining the possible impact of study methodology on the results.

2. Methods

2.1. Study identification and selection

Publications satisfying the following criteria were eligible for consideration:

1. Inclusion of patients with cancer.
2. Investigation of the relationship between Ki-67 expression measured by immunohistochemistry and FLT uptake measured with PET scanning.
3. Publication as a full paper in peer-reviewed scientific journal.

All cancer types were eligible for inclusion and no restrictions were placed on patient characteristics or study design. Studies could include patients with benign conditions, as long as the majority of patients in the sample had cancer. Studies that analysed post treatment changes only were excluded because previous treatment, especially with radiotherapy, is known to affect the reliability of Ki-67 expression.31

2.2. Search methods

A search of studies published between January 1998 and November 2011 in MEDLINE (1948 to November week 4, 2011, Ovid interface), EMBASE (1980–2011 week 48, Ovid interface) and the Cochrane Library was performed. Both subject headings and free text were used for the search. The search was performed with a combination of terms related to PET, FLT and Ki-67. The full electronic search strategy for EMBASE is listed in Table 1. It was performed with no language restrictions and limited to human studies. When results on the same dataset were reported in several publications, only the most recent or complete publication was included in the analysis, in order to avoid overlapping cohorts. The full publications were then obtained and reviewed. Previous reviews on the subject, references reported in the identified studies and the archive indexes of the journal that contained the greatest number of relevant papers were used also in the search.

2.3. Data extraction and management

Data collected from the individual studies covered: (1) overall study characteristics, (2) technical characteristics of both the Ki-67 and the FLT-PET measurement, and (3) the degree of correlation between FLT-PET uptake and Ki-67 score. Details of the included studies are listed in Tables 2 and 3.

The decision as to which technical characteristics of the Ki-67 and FLT PET measurements to include in the data extraction was based on two criteria: (1) aspects important for accurate quantification and (2) information required in order to permit replication of the two tests as per the modified Quality Assessment of Studies of Diagnostic Accuracy (QUADAS) tool.32

For Ki-67 these technical characteristics were the antibody used, the nature of the sample (use of whole surgical sample or biopsy), the method for scoring (manual by a pathologist or automated software), and how the labelling index (LI) was calculated (either measuring the highest proliferation regions or providing an average measurement). When authors reported that the measurements provided were based on an area with the highest proliferation rate this was reported as Ki-67\textsuperscript{max}. If measurements attempted to take into account regional heterogeneity by scoring whole tumour sections, randomly selecting which areas to score, or used different tumour sections, the results were reported as Ki-67\textsuperscript{mean}.

For FLT-PET the data collected covered technical characteristics of the scanner, administered radioactivity, patient preparation, emission scan time, the method used to reconstruct the image, the uptake period after injection of the tracer, and the delineation method used to define the tumour volume. When authors measured the highest FLT uptake this was recorded as SUV\textsubscript{max}. SUV\textsubscript{mean} is the measurement of the average uptake across the whole tumour. When feasible, authors of the studies were contacted by email in order to retrieve missing information on study methods or results. RevMan 5.1 was used for data collection and management.

2.4. QUADAS assessment of methodological quality

The QUADAS tool was selected to assess the methodological quality of the selected studies. QUADAS consists of a scheme of 14 items, formed as questions, for assessing different categories of bias in a study. It has been previously developed and evaluated by Whiting et al.32,33 and described by Hollingworth et al.34 and it is included in the Cochrane Handbook for Diagnostic Test Accuracy Reviews.35 QUADAS-2, a revised version of QUADAS, has been recently published,36 but, as this has not yet been implemented into the
The first step for using the QUADAS tool is to tailor the guidelines for scoring each individual item so that they can be applied to the particular study in question. This is provided as an instruction from the QUADAS developers as they acknowledge that it is not possible to provide a generic description of each item applicable to all reviews. In order to ensure that all reviewers were clear on how studies should be scored for each item, the original QUADAS list was reviewed and a consensus was agreed amongst the authors regarding the guidelines for scoring each item, with a view to making them applicable to studies assessing the correlation between Ki-67 (designated the ‘reference test’) and FLT-PET (designated the ‘index test’) measurements. This process resulted in three items being removed from the list altogether. The QUADAS tool is not intended to be used for weighting data for meta-analysis, and the meta-analysis below was performed entirely independently of the QUADAS assessment. Table 4 shows the original and modified criteria. Each of the 11 items in the modified list was scored as yes (positive) if the publication covered all aspects of the item as described in the modified criteria, no (negative) if it did not, and unclear if no conclusion could be drawn from the information provided.

2.5. Meta-analysis of FLT/Ki-67 correlation

An estimate of the pooled correlation coefficient between Ki-67 LI and SUV FLT uptake was calculated by combining the correlation coefficients obtained in the individual studies. In order to ensure that an unbiased estimate was calculated, the following procedures were performed. Correlation coefficient (r) values were extracted from each study and 95% confidence intervals (CIs) were calculated after applying Fisher’s z transformation. In cases where r values were not reported, these were calculated from the scatter plot graph of FLT uptake versus Ki-67 score. Heterogeneity of the r values between studies was determined by performing Cochran’s Q test and the I² index. The random-effect model (DerSimonian and Laird) was used in place of a fixed-effect model for the pooled analysis. With the random effect model studies are weighted by the inverse of their variance with tau-squared, taking into account the within study variance for estimating the correlation coefficient in each study and the between studies variance (because for example of different methodologies used or possible biological reasons). As a result study weights are assigned with the goal of minimising both sources of variance.

For subgroup analysis, studies were stratified by: (a) the method used to measure Ki-67 LI (average or maximum expression), (b) the use of whole specimen or biopsy samples, and (c) cancer type. Funnel plot, Trim and Fill analysis and Egger’s test were used to investigate publication bias. MIX2 PRO version 2.01-4 and Analyse-it version 2.26 Excel12+ were used for all statistical analyses. For all tests p < 0.05 was considered significant.

3. Results

3.1. Study identification and selection

The original search in EMBASE and MEDLINE identified 159 articles. Searching through the Cochrane database did not identify any articles. In addition, one further study was identified through the reference list of Yap et al. and one was identified through the archive index of the Journal of Nuclear Medicine. After removing duplicates, 94 abstracts were screened according to the evaluation criteria, and 36 in total were selected to be read in full as potentially eligible. 27 prospective studies were selected for inclusion in the review, while nine studies were being excluded for the reasons listed in Fig. 1. Fig. 1 describes the study identification process and results according to the PRISMA guidelines for reporting systematic reviews.

3.2. Study characteristics

The selected studies were published between 2003 and 2011. The median number of patients per study was 20 (range 6–66) with a total number of 608 patients. Studies covered a range of cancer sites as summarised in Table 5.
The most studied tumour group was the brain with six studies. The next well-studied tumour type was lung cancer with five studies in total. Two of these included only lung cancer patients and three investigated other types of cancer as well.
gated thoracic tumours in general. Three groups studied head and neck cancer. The list is completed by breast cancer, gastric cancer, oesophageal, colorectal, lymphoma, sarcoma, hepatocellular, and ovarian cancer.

FLT-PET scanning had variations in acquisition and processing parameters and methods including the injected activity, post injection scanning time and scanning acquisition time. Five studies used dynamic acquisition. Almost all studies provided SUV\textsubscript{max} measurements with half of these also reporting SUV\textsubscript{mean} values. Most groups used manual delineation of the region of interest. Automated delineation using a fixed threshold was the next most common technique with threshold varying between 50% and 80%.

For scoring of Ki-67 expression, the majority of studies measured the highest Ki-67 expression (Ki-67\textsubscript{max}) in the sample, which was usually obtained from a biopsy, and the counting was performed by eye rather than image analysis software. MIB-1 antibody from Dako was used for the majority of the studies (Table 3).

### 3.3. QUADAS assessment of methodological quality

Based on QUADAS scores, studies were ranked for each item as presented in Fig. 2. None of the studies scored ‘YES’ for all 11 items. Items two and three scored the lowest – lack of a clear description of the selection criteria being commonest finding. Specifically most authors did not report a detailed description of the inclusion criteria or a specific time period for recruitment. For item three (acceptable delay between reference and index tests) only a third of the studies scored positively. Most studies either provided no information about the time interval between the execution of the biopsy or surgical resection and the FLT scan or, if

<table>
<thead>
<tr>
<th>Author</th>
<th>Cancer</th>
<th>No</th>
<th>Sample</th>
<th>Antibody</th>
<th>Method</th>
<th>Image analysis</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buck et al. (2003)</td>
<td>Lung</td>
<td>20</td>
<td>Both</td>
<td>Dianova</td>
<td>Ki-67\textsubscript{mean}</td>
<td>Manual</td>
<td>0.92</td>
</tr>
<tr>
<td>Dittmann et al. (2003)</td>
<td>Various</td>
<td>9</td>
<td>Both</td>
<td>Dako</td>
<td>NI</td>
<td>NI</td>
<td>0.22</td>
</tr>
<tr>
<td>Francis et al. (2003)</td>
<td>Colorectal</td>
<td>10</td>
<td>Surgery</td>
<td>Dako</td>
<td>Ki-67\textsubscript{mean}</td>
<td>Manual</td>
<td>0.8</td>
</tr>
<tr>
<td>Chen et al. (2005)</td>
<td>Brain</td>
<td>14</td>
<td>Surgery</td>
<td>Dako</td>
<td>NI</td>
<td>Manual</td>
<td>0.84</td>
</tr>
<tr>
<td>Choi et al. (2005)</td>
<td>Brain</td>
<td>9</td>
<td>Surgery</td>
<td>MIB-1</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>0.82</td>
</tr>
<tr>
<td>Cobben et al. (2004)</td>
<td>Sarcoma</td>
<td>19</td>
<td>Surgery</td>
<td>Immunotech</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>0.652</td>
</tr>
<tr>
<td>Smyczek et al. (2004)</td>
<td>Breast</td>
<td>12</td>
<td>Biopsy</td>
<td>NI</td>
<td>NI</td>
<td>Semi-automatic</td>
<td>0.14</td>
</tr>
<tr>
<td>Kenny et al. (2005)</td>
<td>Breast</td>
<td>12</td>
<td>Biopsy</td>
<td>Novocastra</td>
<td>Ki-67\textsubscript{mean}</td>
<td>Manual</td>
<td>0.79</td>
</tr>
<tr>
<td>Westreenen et al. (2005)</td>
<td>Oesophageal</td>
<td>10</td>
<td>Both</td>
<td>Dako</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>−0.76</td>
</tr>
<tr>
<td>Buck et al. (2006)</td>
<td>Lymphoma</td>
<td>21</td>
<td>Biopsy</td>
<td>Dianova</td>
<td>Ki-67\textsubscript{mean}</td>
<td>Manual</td>
<td>0.84</td>
</tr>
<tr>
<td>Saga et al. (2006)</td>
<td>Brain/lymphoma</td>
<td>20</td>
<td>Surgery</td>
<td>Dako</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>0.84</td>
</tr>
<tr>
<td>Yap et al. (2006)</td>
<td>Lung</td>
<td>15</td>
<td>Both</td>
<td>Dako</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>0.6</td>
</tr>
<tr>
<td>Yamamoto et al. (2007)</td>
<td>Lung</td>
<td>18</td>
<td>Surgery</td>
<td>Dako</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>0.77</td>
</tr>
<tr>
<td>Hatakayama et al. (2008)</td>
<td>Brain</td>
<td>18</td>
<td>Surgery</td>
<td>Novocastra</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>0.89</td>
</tr>
<tr>
<td>Linecker et al. (2008)</td>
<td>Head and Neck</td>
<td>18</td>
<td>Biopsy</td>
<td>Dako</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>0.058</td>
</tr>
<tr>
<td>Ullrich et al. (2008)</td>
<td>Brain</td>
<td>11</td>
<td>Image guided biopsy and surgery</td>
<td>DCS</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>0.57</td>
</tr>
<tr>
<td>Eckel et al. (2009)</td>
<td>Hepatocellular</td>
<td>12</td>
<td>Both</td>
<td>Dako</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>0.66</td>
</tr>
<tr>
<td>Kameyama et al. (2009)</td>
<td>Gastric</td>
<td>20</td>
<td>Both</td>
<td>Dako</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>−0.32</td>
</tr>
<tr>
<td>Price et al. (2009)</td>
<td>Brain</td>
<td>35</td>
<td>Image guided biopsy</td>
<td>Dako</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>0.66</td>
</tr>
<tr>
<td>Yamamoto et al. (2009)</td>
<td>Colorectal</td>
<td>26</td>
<td>Surgery</td>
<td>Dako</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>0.04</td>
</tr>
<tr>
<td>Troost et al. (2010)</td>
<td>Head and Neck</td>
<td>20</td>
<td>Surgery</td>
<td>Abcam</td>
<td>Ki-67\textsubscript{mean}</td>
<td>Automatic</td>
<td>0.044</td>
</tr>
<tr>
<td>Brockenbrough et al. (2011)</td>
<td>Lung</td>
<td>22</td>
<td>Surgery</td>
<td>Dako</td>
<td>Ki-67\textsubscript{max} and mean</td>
<td>Manual</td>
<td>0.57, 0.69</td>
</tr>
<tr>
<td>Contractor et al. (2011)</td>
<td>Breast</td>
<td>9</td>
<td>Biopsy</td>
<td>Novocastra</td>
<td>Ki-67\textsubscript{mean}</td>
<td>Manual</td>
<td>0.82</td>
</tr>
<tr>
<td>Herrmann et al. (2011)</td>
<td>Lymphoma</td>
<td>43</td>
<td>Biopsy</td>
<td>Dianova</td>
<td>Ki-67\textsubscript{mean}</td>
<td>Manual</td>
<td>0.21</td>
</tr>
<tr>
<td>Hoshikawa et al. (2011)</td>
<td>Head and Neck</td>
<td>31</td>
<td>Both</td>
<td>Dako</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>−0.10</td>
</tr>
<tr>
<td>Ott et al. (2011)</td>
<td>Gastric</td>
<td>37</td>
<td>Biopsy</td>
<td>Dako</td>
<td>Ki-67\textsubscript{max}</td>
<td>Manual</td>
<td>0.017</td>
</tr>
<tr>
<td>Richard et al. (2011)</td>
<td>Ovarian</td>
<td>6</td>
<td>Surgery</td>
<td>Ventana</td>
<td>Ki-67\textsubscript{mean}</td>
<td>Manual</td>
<td>0.77</td>
</tr>
</tbody>
</table>

\(\text{Ki-67\textsubscript{max}} = \) measurement of Ki-67 highest scoring, \(\text{Ki-67\textsubscript{mean}} = \) measurement of Ki-67 average scoring.

\(^a\) Number of patients who had both FLT and Ki-67 measurements based on which the correlation was calculated.

\(^b\) Correlation with Ki-67\textsubscript{mean} (\(r = 0.57\)) and with Ki-67\textsubscript{max} (0.69).
1. Was the spectrum of patient’s representative of the patient's who will receive the test in practice? Representitive spectrum?
Modified QUADAS: Participants in the study should be patients with cancer selected following a prospective patient inclusion. As previous treatment could affect scoring of Ki-67, we only considered as representative, studies that included patients without any previous treatment or if the time interval between last treatment and sample selection was more than 3 months. Details should have been provided about age distribution, female to male ratio and disease description as minimum information (patient characteristics table).

2. Were selection criteria clearly described?
Modified QUADAS: The description of the selection criteria was considered sufficient when authors reported a specific clear definition of inclusion and exclusion criteria for entry into the study and a time period and location of recruitment.

3. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?
Modified QUADAS: A time period of 4 weeks was chosen as a cut off point; studies that performed the biopsy/surgery and FLT scan within 4 weeks for more than 80% of the participants, scored this item as Yes. When no information was given or when the time of either the biopsy/surgery or the PET scan only was specified this item was scored as unclear.

4. Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis? Partial verification avoided?
Modified QUADAS: Although in the original QUADAS this item refers to the participants sample and represents partial verification bias, in our review it refers to tissue sample. As all studies included were validation studies aiming to compare quantitative values obtained from the immunohistochemistry labelling index of Ki-67 (reference standard) and FLT-PET uptake (index test) this item was not applicable in its original form as all studies were free from verification bias. If Ki-67 score was performed on the whole specimen (either as highest or average value) or a biopsy sample but only as average measurement, then this item was scored as yes. If a study has used only biopsy material and scored only the area of highest mitotic activity then this item was scored as No. If no information was given by the authors, this was scored as unclear.

5. Did patients receive the same reference standard regardless of the index test result? Differential verification avoided?
Modified QUADAS: If patients have Ki-67 measurements regardless of the FLT uptake then score this item as Yes.

6. Was the execution of the index test described in sufficient detail to permit replication of the test?
Modified QUADAS: To score as Yes the description of this item should include: Injected dose, uptake period, emission time, reconstruction method, patient preparation, volume of interest delineation method. Method of delineation is mandatory in order for studies that report FLT mean SUV values to score Yes. If they are only reporting SUVmax values then this item can score as Yes even without detail description of delineation method.

7. Was the execution of the reference standard described in sufficient detail to permit its replication?
Modified QUADAS: To score as Yes the description of this item should as minimum include: Antibody used, antigen retrieval, dilution, incubation time, visualisation method and scoring method. To score as Yes the description of this item should include: Injected dose, uptake period, emission time, reconstruction method, patient preparation, volume of interest delineation method. Method of delineation is mandatory in order for studies that report FLT mean SUV values to score Yes. If they are only reporting SUVmax values then this item can score as Yes even without detail description of delineation method.

8. Were the index test results interpreted without knowledge of the results of the reference standard? Index test results blinded?
Modified QUADAS: To confirm that authors were blinded to the results of the reference standard then a clear statement in the paper had to be given. If it is clear that no blinding occurred then score as No. If no statement on blinding was given the item was scored unclear.

9. Were the reference standard results interpreted without knowledge of the results of the index test? Reference test results blinded?
Modified QUADAS: To confirm that authors were blinded to the results of the index test when interpreting the reference test a clear statement in the text had to be given. If it was clear that no blinding occurred score as No. If no statement on blinding was given the item was scored unclear.

10. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? Relevant clinical information?
Modified QUADAS: Both SUV and Ki-67 measurements are quantitative assays and as such clinical data should not be available during the interpretation of tests results. If authors report that this is the case score this item as Yes. If clinical data were available during the interpretation of the FLT scan score this item as No.

11. Were withdrawals/missing values from the study explained? Withdrawals explained?
Modified QUADAS: If for any patients in the study either Ki-67 score or FLT uptake was not reported but authors provided a valid explanation then score this item as Yes. Otherwise score this item as No.

12. Is the reference standard likely to correctly classify the target condition?
Modified QUADAS: The reference test in our case is Ki-67 scoring. As Ki-67 is the clinical gold standard for assessing proliferation status in tissue and because ki-67 scoring was one of the selection criteria for including studies in the review this item was removed from the checklist.

13. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?
Modified QUADAS: This item was removed from the list as the two tests are not related.

14. Were uninterpretable/intermediate test results reported?
Modified QUADAS: As these are quantitative measurements and not binary test outcomes (positive/negative) this item was removed from the list.
documented, the interval was more than 4 weeks. Item four refers to the use of the whole surgical sample to derive the Ki-67 score or, if only biopsies were used, the use of Ki-67\textsubscript{mean} values rather than Ki-67\textsubscript{max} values (20 studies scored positive, five negative and two were unclear).

In the majority of studies Ki-67 labelling index was measured in the area (or areas) of the tumour with the highest expression. Four studies that provided no information\textsuperscript{13,18,20,21} were assigned to the Ki-67\textsubscript{max} subgroup based on the assumption that this is what the pathologist would score in routine clinical practice.

Items six and seven (description of the execution of index and reference test respectively) scored relatively well. All studies provided information about equipment characteristics and protocols for both the scanning and histopathology procedures, but the amount of information varied. For Ki-67 immunohistochemistry most authors chose to refer directly to the manufacturer’s instructions, rather than providing a detailed description of the staining or scoring processes. In general the FLT PET scan protocol was well described.

Assessing Ki-67 LI or FLT-PET uptake in a blinded fashion, availability of clinical data during interpretation of FLT-PET and unexplained missing values of the Ki-67 scoring results were items which most studies did not report. Only approximately half of the studies stated that interpretation of the test results was done in a blinded fashion or that other clinical results were not available.
In twenty studies withdrawals were explained and in seven studies they were not.

3.4. Meta-analysis of FLT/Ki-67 correlation

All studies provided data that was suitable for meta-analysis. Meta-analyses were based on the correlation between Ki-67 score and FLT-PET SUV values. PET kinetic parameters were not used as all dynamic studies, apart from one, that reported a strong correlation between kinetic parameters and Ki-67 also reported a strong correlation between SUV and Ki-67, albeit with lower values. For four studies, the \( r^2 \) values were calculated based on the graphical representation to determine the sign. For one study, \( r \) was calculated directly from the graph of SUV max versus Ki-67 measurements provided in the paper. The pooled \( r \) for all studies was 0.55 (95% CI 0.37–0.69, \( p < 0.001 \)) and was highly heterogeneous \( I^2 = 82\% \) (\( Q = 147, p < 0.001 \)).

For subgroup analysis based on Ki-67 scoring and nature of the sample studies were divided in different groups, based on the use of Ki-67 max or Ki-67 mean measurements and whether they use whole surgical specimens or biopsy samples.

Nine studies that used either biopsy or surgery were included in the Ki-67 max group with \( r = 0.70 \) (95% CI = 0.43–0.86, \( p < 0.001 \)) and heterogeneity \( I^2 = 81\% \) (\( Q = 43, p < 0.001 \)). When studies were further divided (data not shown) into biopsy versus surgery subgroups pooled correlation of the Ki-67 max/biopsy subgroup was identical with the one provided by the Ki-67 mean/surgery subgroup (\( r = 0.70, 0.26–0.90, p < 0.01 \) versus \( r = 0.70, 0.23–0.90, p < 0.01 \)). Both subgroups had similar heterogeneity (\( I^2 = 82\% \), \( Q = 16, p < 0.001 \) versus \( I^2 = 84\%, Q = 25\%, p < 0.001 \)). For the sensitivity analysis removal of Troost et al. and Herrmann et al. (the only outliers) greatly reduced heterogeneity in the subgroup without significantly affecting the \( r \) value (\( r = 0.81, 95\% \) CI = 0.69–0.88, \( p < 0.001 \), \( Q = 9, I^2 = 36\% \), \( p = 0.15 \)) (Fig. 3).

Studies that have provided Ki-67 max measurements were divided into two subgroups based on the nature of the sample. Nine studies that used whole surgical specimen and seven studies that used biopsy for either all or more than one third of their samples were independently analysed. Studies where this was unclear were excluded from the analysis. The study of Yamamoto et al. was the only outlier in the surgery group, this is the only study in this group with colorectal cancer while the others are studies in lung, brain cancer and sarcoma. Removal of this study...
resulted in $F = 2\%$ ($Q = 7, p = 0.41$). The pooled correlation for studies that used non-image guided biopsy was not statistically significant. ($r = 0.04,$ 95% CI = −0.18–0.26, $p = 0.71$). Heterogeneity for the biopsy subgroup was not significant $F = 27\%$ ($Q = 8, p = 0.21$) (Fig. 3).

The studies from Price et al.18 and Ullrich et al.14 that have used image guided biopsy were analysed as an independent subgroup with $r = 0.64$ (95% CI = 0.43–0.79, $p < 0.001$) and $F = 0\%$ ($Q = 0, p = 0.71$) (Fig. 3).

Results for subgroup analysis based on cancer type are shown in Fig. 4. Six studies in brain cancer9,10,12–14,18 resulted in $r = 0.78$, (95% CI = 0.67–0.86, $p < 0.001$), with $F = 22$ ($Q = 6, p > 0.05$). The pooled $r$ for four studies in lung cancer5,8,16,43 was 0.78 (95% CI = 0.56–0.89, $p < 0.001$), with $F = 69\%$ ($Q = 9, p < 0.05$). For breast cancer three studies were identified17,20,44 providing an $r = 0.67$, (95% CI = 0.20–0.89, $p < 0.01$) with $F = 70\%$ ($Q = 6, p < 0.05$). In head and neck, three studies were identified22,25,26 providing an $r = -0.01$, (95% CI = −0.26–0.22, $p > 0.05$) with $F = 0$ ($Q = 0, p = 0.84$).

Funnel plot and Trim and Fill analysis did not reveal any publication bias (Egger’s test = 1.7, $p > 0.05$).

4. Discussion

Our systematic review of published studies investigating the correlation between Ki-67 expression and FLT-PET uptake was designed with two objectives. First, to investigate the variations in methods used in these studies and second, to perform a meta-analysis of the reported correlation values examining the impact of study methods on the results. To the best of our knowledge this is the first meta-analysis applied to studies addressing the correlation between an imaging and a biological biomarker. The main findings from our two analyses are: (a) in the studies we examined there are variations in several aspects of the methods and reporting and (b) given an appropriate study design the FLT/Ki-67 correlation is significant and independent of cancer type.

The assessment of the methodological quality revealed potential sources of bias in the studies reviewed. In most studies the time between the FLT-PET scan and surgery or biopsy was not clearly defined highlighting the fact that this is as much an outcome of inadequate reporting as of trial design. Likewise, the study inclusion criteria and the specific time interval of the recruitment period are not adequately described. Whether reporting was done in a blinded fashion and whether there was access to further clinical data when reporting Ki-67 or FLT-PET measurements were also under-reported in the majority of studies.

Our statistical analysis indicates that the study design, and more specifically the method of Ki-67 expression measurement, has a direct effect on the $r$ values between Ki-67 expression and FLT uptake. Studies that have used whole surgical samples and Ki-67 mean measurements achieved higher $r$ values. Ki-$67_{\text{max}}$ measurements especially from a biopsy sample are likely to have reduced reproducibility as they do not take into account the degree of intratumour heterogeneity for this marker. Recently Goodell et al.55 showed that different methods for determination of the proliferative index yielded different results, with inconsistent grading in two thirds (67%) of the tumours examined and that Ki-$67_{\text{max}}$ measurements may overestimate in some cases the proliferative index. If whole tumour specimens are used then an experienced pathologist can often identify accurately these areas in the specimen (in the invasive front for example). Average immunohistochemistry measurements are more reproducible, stable, and less affected from bias.56 They are also more similar to the way FLT-PET can produce an integrated average value and therefore more likely to correlate with the SUV$_{\text{mean}}$. As assessing the whole surgical sample with a tissue based technique is not feasible, not only in routine clinical practice but also in a research setting, future studies would benefit from the use of PET guided biopsies. This introduces its own challenges in particular to ensure accurate image registration, especially in parts of the body that are affected by motion such as lung.57

Whilst the whole volume of a PET scan is available to identify the SUV$_{\text{max}}$, in tissue, immunohistochemistry usually assesses only a few micrometre thick sections of 1–6 square centimetres. But a 3 μm-thick tissue section of $3 \times 2$ cm, represents only $1/19,000$ of the volume of a tumour with a diameter of 4 cm.58 In tumours with high heterogeneity the concordance between biopsy and whole tumour might be low but in tumours with relative homogeneity of the tissue biomarker distribution then even a biopsy sample might be sufficient to score the tissue biomarker component. Fernebro et al. examined the expression of Ki-67 in surgical specimens from sarcoma tumours. They demonstrated a higher fraction of Ki-67 positivity in the tumour periphery compared to the tumour centre but the relative difference was less than 10%.59 On the contrary Muller et al. provided evidence of large differences owing to strong intratumour heterogeneity of Ki-67 expression when three different proliferation indices (maximum, average or located at the tumour invasion front) were determined in gastric cancer.60 Intratumoural heterogeneity could also vary depending on the biomarker under investigation. Moertel et al.61 in their recent study on head and neck tumours concluded that representative biopsies of Ki-67 and EGFR staining varied significantly depending which parts of the tumours were examined and that for translational research purposes whole tumour sections should be examined in order to get accurate staining measurements. Toles et al. propose that, especially for new quantitative biomarkers, therapeutic response
is dictated by the level of expression rather than simple cut off points, and the optimal size of the tissue sample must be determined on a marker-by-marker basis. In addition to the above considerations it is interesting to note that there is increasing evidence regarding the prognostic or predictive value of the heterogeneity of imaging-based biomarkers.

Variations in studies results cannot only be attributed to biology. They can also be introduced by the way data are generated. Examining the technical characteristics of FLT-PET scanning in the reviewed studies, it is clear that there is no consensus on the way they should be performed. SUV\textsubscript{max} is easy to measure with available commercial software and it is not subject to inter-observer variability as it is not dependent on the delineation method used. It is however affected by the statistical properties of the image, the voxel size and tumour motion. A possible source of measurement error is associated with the size of the tumour lesion with small tumours been affected the most by the partial volume effect. The magnitude of this error will however vary amongst different studies. For example in Vesselle et al.\textsuperscript{47} although partial volume correction improved the correlation coefficient between FLT and Ki-67 it did so by only 0.05 (non-corrected $r = 0.78$ versus corrected $r = 0.83$).

Specific guidelines for performing and reporting standard clinical FDG scans exist.\textsuperscript{67} Similar guidelines should be developed and applied to validation studies of novel PET tracers in order to generate more robust, comparable data. To this end initiatives like the American College of Radiology imaging Network (ACRIN) and the NCRI PET Research Network, amongst others, would be most useful in addressing the standardization
barriers to conducting early and late phase clinical trials with PET.\textsuperscript{68,69} Similar efforts have been addressed for immunohistochemistry scoring. In order to minimise bias and increase reproducibility for such cases, the EORTC-GCCG published a consensus paper in 1998 on immunohistochemical staining and scoring based on a systematic random sampling protocol, aiming to achieve the most representative sample and the highest reproducibility.\textsuperscript{56} Recently the Breast Cancer Working Group published recommendations on pre-analytical and analytical assessment, and interpretation and scoring of Ki-67 in breast cancer based on current evidence.\textsuperscript{70} It must be recognised that Ki-67 is selected as an excellent marker based on its biological expression assessed.
It is difficult to apply in a systematic way beyond a laboratory experiment setting using directly comparable test and controls. Use in vivo is associated with variability caused by sampling and counting protocols. Novel assessments of relative scoring combined with spatial assessment may be a better way to validate imaging studies than simple sample counting.

Our study has a few potential limitations. One of the main issues is that the number of recruited patients in respective studies is rather small in some cases. However, the absence of publication bias and the analysis of 27 different studies with a total of 509 patients included strengthen our results. In addition, it could be argued that because of the variations in the way FLT-PET and Ki-67 immunohistochemistry was performed pooling of the data was inappropriate. However 70% of the studies were performed by the same scanner model (ECAT HR+ Siemens PET), 75% used a post injection time between 40–60 min and 60% of them performed manual Ki-67 scoring with the use of the same antibody (Dako) under the manufacturer’s guidelines. Moreover, the studies incorporated into this analysis were from institutions with high levels of expertise in the conduct of such trials so that despite the variations in protocols high quality was ensured in most cases. The most important bias in the studies included is that while with FLT-PET spatial information is acquired across the entire tumour, Ki-67 spatial information is limited in some occasions to only a small biopsy sample which might not be representative of the whole tumour and so may compromise the accuracy of the measurement. By sub-grouping studies based on the method used for Ki-67 scoring, the nature of the tissue sample and the cancer type we are not eliminating all sources of variance completely, as this is not feasible for any kind of meta-analysis, but it does allow us to identify the biggest sources of bias in these studies. Therefore, we believe that the use of random-effects modelling and sensitivity analyses address this heterogeneity adequately.

Studies investigating the correlation between imaging-based and tissue-based biomarkers involve procedures that are time-intensive, methodologically challenging and expensive. When a large sample size is needed or when investigating new tracers, it is essential to perform these studies in a way that would allow cross validation and pooled analysis of the results. Despite the variations recorded above our subgroup analysis indicated that the correlation is significant and independent of cancer type provided an appropriate study design.

5. Authors’ conclusions

In validation studies comparing FLT uptake with Ki-67 expression the correlation is affected by methods used and study design. The correlation is significant and independent of cancer type provided a study design with either the use of Ki-67 average measurements, regardless of nature of sample, or whole surgical samples when measuring Ki-67 maximum expression. Sufficient data to support a strong correlation for brain, lung and breast cancer exist. However, larger, prospective studies with improved study design are warranted to validate these findings for the rest of the cancer types. It is suggested that future validation studies of PET imaging biomarkers will benefit from the following:

- The use of mean and maximum measurements for both the SUV and Ki-67 scoring. The use of \( \text{SUV}_{\text{peak}} \) should also be considered in future studies, this uses a 1 ml volume containing the highest values of SUV.
- When studies use only biopsy samples then the use of PET guided biopsy particularly for highly heterogeneous tissues would be beneficial. Alternatively in the case of whole surgical specimens the use of extensive sampling and assessment for Ki-67 staining to ensure the full range and maximum expression in the whole tumour are correctly identified.
- The investigation of tumour heterogeneity both in tissue and in image with the development of appropriate software analysis tools.
- The development of reporting guidelines that will also include a detailed report of the method and execution of the PET and the histopathology measurements.

Conflict of interest statement

None declared.

Funding

None.

Acknowledgements

We would like to thank Clare Crowley, Information Specialist from Kings College London, for her help during database search. We would also like to acknowledge the valuable guidance of Dr. Toby Prevost from the department of Primary Care and Public Health Sciences, King’s College London during the development of the meta-analysis.

We would also like to acknowledge King’s College London and UCL Comprehensive Cancer Imaging Centre. Funded by the CRUK and EPSRC in association with the MRC and DoH (England).

References


学霸图书馆
www.xuebalib.com

本文献由“学霸图书馆-文献云下载”收集自网络，仅供学习交流使用。

学霸图书馆（www.xuebalib.com）是一个“整合众多图书馆数据库资源，提供一站式文献检索和下载服务”的24小时在线不限IP图书馆。
图书馆致力于便利、促进学习与科研，提供最强文献下载服务。

图书馆导航：
图书馆首页 文献云下载 图书馆入口 外文数据库大全 疑难文献辅助工具