Primary Tumor Standardized Uptake Value Measured on Fluorodeoxyglucose Positron Emission Tomography Is of Prognostic Value for Survival in Non-small Cell Lung Cancer

Update of a Systematic Review and Meta-Analysis by the European Lung Cancer Working Party for the International Association for the Study of Lung Cancer Staging Project

Marianne Paesmans, MSc, Thierry Berghmans, MD, PhD,† Michele Dusart, MD,‡ Camillo Garcia, MD,§ Claude Hossein-Foucher, MD, PhD,‖ Jean-Jacques Lafitte, MD, PhD,¶ Céline Mascaux, MD, PhD,# Anne-Pascale Meert, MD, PhD,† Martine Roelandts, MD,** Arnaud Scherpereel, MD, PhD,¶ Vanessa Terrones Munoz, MD,† and Jean-Paul Sculier, MD, PhD†; for the European Lung Cancer Working Party, and on behalf of the IASLC Lung Cancer Staging Project*

Introduction: Few validated prognostic factors are available for survival in patients with lung cancer. [^{18}F]-fluoro-2-deoxy-D-glucose positron emission tomography has been shown to be of additional value to conventional imaging for staging lung cancer. The prognostic value of this lung tumor metabolic activity was studied in a first systematic review of studies published until 2006.

Methods: As further studies have appeared since 2006, this report has as objective to confirm and to estimate with less variability the prognostic value of primary tumor standardized uptake value (SUV) measured with [^{18}F]-fluoro-2-deoxy-D-glucose positron emission tomography on the basis of an updated search of eligible studies.

Results: Ten additional studies were eligible for the updated review and eight of them provided, in the publication, data allowing survival results aggregation. All together, 21 studies were analyzed. Comparing patients with low and high SUV, using preferentially the median SUV value of each study as threshold, we obtained a poor prognostic value for high SUV compared with low SUV with an overall combined hazard ratio of 2.08, significantly different from one with a 95% confidence interval ranging from 1.69 to 2.56. No interaction between older and newer studies was detectable (P = 0.60) as well as between studies having selected non metastatic patients or studies without selection criterion related to stage (P = 0.46).

Conclusions: We confirmed the results of our previous review showing that SUV is potentially a very interesting factor for predicting patient outcome. We believe that a meta-analysis based on individual patient data would be of great value as allowing to assess the independent prognostic value, to take into account some factors responsible for heterogeneity between studies (SUV assessment method, disease stage, and histology), and to update survival data. We are planning to conduct such a meta-analysis on behalf of the International Association for the Study of Lung Cancer Staging Project.

Key Words: Lung cancer, ^{18}FDG-PET scan, Prognostic factor, Survival, Systematic review.

(*J Thorac Oncol.* 2010;5: 612–619)

Up to now, there are, for lung cancer, very few validated prognostic factors in use in clinical practice. Although many clinical, pathologic, routine laboratory markers, molecular biologic markers, and gene signatures have been suggested as possible univariate or independent prognostic factors, only two characteristics are definitely established as independently associated with prognosis: performance status and disease stage.[1] This has been recently confirmed by the International Association for the Study of Lung Cancer (IASLC) Staging Project. Within the context of this project,

*Data Centre, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium; †Department of Intensive Care and Thoracic Oncology, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium; ‡Medical Direction, Hôpital de la Citadelle, Liege, Belgium; §Nuclear Medicine Department, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium; ‖Nuclear Medicine Department and ¶Pneumology Department, CHU Lille and University of Lille II, Lille, France; #National Fund of Scientific Research (F.R.S.–FNRS), Brussels, Belgium; and **Radiotherapy Department, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium.

Disclosure: The authors declare no conflicts of interest.

Address for correspondence: M. Paesmans, Data Centre, Institut Jules Bordet, Université Libre de Bruxelles, Rue Héger-Bordet 1, B–1000 Bruxelles, Belgium. E-mail: marianne.paesmans@bordet.be

J. P. Sculier is member of the International Staging Committee of the International Association for the Study of Lung Cancer (IASLC).

ISSN: 1556-0864/10/0505-0612

data on more than 100,000 patients have been collected and used for proposing a revision of the 6th edition of the tumor, node, metastasis classification for lung cancer but also for looking at prognostic factors. The study on prognostic factors has however been limited by the small number of covariates being present in the worldwide existing data bases. This is a further illustration that there are very few factors universally accepted as important for predicting outcome of the disease.[2]

Tumor, node, metastasis staging assessment was traditionally based on conventional imaging but since a decade, positron emission tomography (PET) with the glucose analogue 2-[^{18}F]-fluoro-2-deoxy-D-glucose (^{18}FDG) has been successfully introduced in the staging procedures. Indeed, several meta-analyses have shown that it improves the accuracy of the staging assessment.[3] This has motivated our group for performing a systematic review and meta-analysis of the literature aiming to analyze the possible prognostic value of FDG-PET for survival.[4] The results of this review were published in January 2008 and suggested that a high standardized uptake value (SUV) might be a poor prognostic factor for survival. The estimated hazard ratio (HR) that we calculated from the combination of 13 studies,[5–17] using a random-effects model, was 2.27 (95% confidence interval [CI], 1.70–3.02). However, because our search for identifying studies eligible for our review (in June 2006) was stopped, several further publications on the same topic occurred, some of them focusing on specific disease stages. We therefore decided that it was worth to update our review.

MATERIALS AND METHODS

For obvious reasons, we followed the same methodology as for our previous report, but we summarized it shortly here after.

We looked for studies dedicated to lung cancer only, performing ^{18}F-FDG-PET assessment on a dedicated device before any treatment and assessing the relationship between pretherapeutic SUV and survival. We only considered full publications (excluding abstracts) in English, French, or Dutch in peer-reviewed journals.

We electronically searched for studies on Medline using keywords specified in our previous report and manually checked the bibliographies of the already selected studies. We stopped the search in January 2009.

Twelve readers (11 physicians and 1 statistician) extracted data from the publications for reporting on the individual characteristics and results for survival of the identified articles as well as data for assessing the methodological characteristics of the reports—on the basis of a published scale. We scored both the clinical and the PET reports (22 and 20 items, respectively) and expressed the results in percentage of the maximal theoretical value that can be obtained. For details about the items that we considered, we refer the reader to our primary review on the same topic.[4]

We descriptively analyzed the scores and looked for any difference between the new studies compared with the older ones.

We cautiously looked at overlaps between cohorts of patients in the different publications. In case of such overlaps, the most recent publication or the one reporting the most accurately the survival relationship between SUV and survival was used. If needed, authors were contacted to confirm or to reject the assumptions of intersections between patients included in several analyses.

For assessing the impact of SUV on survival, we considered, as far as possible, for each study, two groups of patients: one with low SUV (preferably using the median of the observed SUV distribution) and one with high SUV. We thereafter attempted to extract an estimated HR for the comparison of the two survival distributions using as reference the group of patients with low SUV and an estimate of the variance of this estimated HR. Several methods were used depending on the available information in the individual publications as in our previously published meta-analysis[4]: we looked first, for the univariate HR estimate and 95% CI if provided by the author; second, for the logrank test result, number of events, and numbers of patients in each group allowing to retrieve the HR; third, we read the survival rates on the survival curves; and finally, we searched for an adjusted HR (for several baseline factors considered by the authors).

Selection bias was assessed graphically using funnel plot.[18] Heterogeneity between studies was assessed by using a χ^2 test for heterogeneity, and combined HRs were obtained by the use of fixed-effects models in case of absence of heterogeneity and of random-effects models otherwise.[4] Interaction tests were performed using χ^2 tests.

As far as possible, the median values of SUV uptake was used for defining the threshold splitting the patients into two groups. Indeed, the so-called "best cut-off" method has been shown to be associated with a high probability of false positive result and to provide a biased, unreliable, and nonreproducible estimate of the prognostic impact of the tested covariate.[19] We performed sensitivity analyses including and excluding the studies where only results according to the best cut-off were available.

The a priori specified objectives of the analysis were to combine all studies and to look for interactions between publication time (new studies versus older ones), stage and histology, whenever possible. As secondary objective, we planned to assess the possibility of getting aggregated results for disease free survival.

All reported CIs have a confidence level of 95%, and all reported p values are two-tailed. We used 5% as threshold for significance.

RESULTS

Eligible Studies

Thirteen studies, published between 1998 and 2006, were identified and considered eligible from our previous search with a total number of 1474 patients included in the different series.

We selected 14 publications published between 2006 and 2008.[20–33] Three of them were considered as ineligible: one because many patients had treatment before SUV assessment,[27] one[21] because it was uncertain that SUV assessment was done before treatment and one because the authors did

TABLE 1. Characteristics of the Patient Populations Included in the 24 Eligible Studies

Study	Year of Publication	No. of Patients	Type of SUV	ISS	% ADC	% SCC	Stage	% Stage IV
Ahuja et al.[5]	1998	155	Mean	1997	?	?	I–IV	?
Sugawara et al.[6]	1999	38	Max	1986	50	32	I–IV	13
Vansteenkiste et al.[7]	1999	125	Max	1997	25	54	I–IIIB	0
Dhital et al.[8]	2000	77	Max	1986	23	58	?	?
Higashi et al.[9]	2002	57	Mean	1997	67	0	I–III	0
Jeong et al.[10]	2002	73	Max	1997	41	51	I–IV	?
Downey et al.[11]	2004	100	Max	1997	67	24	?	?
Borst et al.[a][12]	2005	51	Max	?	25	33	I–III	0
Cerfolio et al.[b][13]	2005	315	Max	1997	31	51	I–IV	10
Port et al.[14]	2005	64	?	?	88	8	?	?
Sasaki et al.[15]	2005	162	Max	1997	46	43	I–III	?
Eschmann et al.[16]	2006	137	Mean	1997?	29	45	IIIA/IIIB	0
Prevost et al.[17]	2006	120	Mean/max	1997	49	38	I–IV?	?
Raz et al.[23]	2006	36	?	1997	0	0	?	?
de Jong et al.[24]	2007	66	Max	1997	35	53	I–IIIA	0
Downey et al.[25]	2007	487	Max	1997	69	21	I–IV	2
Lee et al.[26]	2007	19	?	1997	?	?	I–IV	26
Na et al.[27]	2007	57	Max	1997	33	53	IIIA N2	0
van Baardwijk et al.[28]	2007	102/46[c]	Max	1997	30	57	I–IIIB	0
Vesselle et al.[29]	2007	208	Max	1997	38	27	I–IV	19
Zhang et al.[30]	2007	82	Max	1997	?	56	I–III	0
Goodgame et al.[31]	2008	136	Max	1997	52	35	I	0
Hanin et al.[32]	2008	96	Max	1997	47	48	I/II	0
Hoang et al.[33]	2008	214	Max	1997	38	30	IIIA/IIIB/IV	43

[a] Study considered as not fully eligible because of inclusion of 11 patients without histological proof of lung cancer but included in the primary review.[4]
[b] Study considered as not fully eligible in the primary review[4] because it was unclear that histological proof of lung cancer was obtained for all patients (but first author confirmed later that it was the case).
[c] Fifty-six of these 102 patients were included in Ref 7.
SUV, standardized uptake value; ISS, International Staging System; ADC, adenocarcinoma; SCC, squamous cell carcinoma.

not measure SUV on the primary tumor but considered the highest SUV value on all the identified lesions.[22]

Thus, in total, 24 studies were eligible for our review, 13 considered as "older" publications from the first systematic review and 11 called the "newer" publications published after our previous meta-analysis. Some characteristics of the patients populations considered in these 24 studies are reported in Table 1. The numbers of patients included in the different studies ranged from 19 to 487. Metastatic patients are under-represented in the studies; half of them explicitly mentioning that inclusion of stage IV patients were not considered. Only one study is dedicated to advanced stages (from IIIA to IV) and, even in this study, the rate of stage IV patients is 43%, rate which is still below the rate of stage IV patients expected in a general population of newly diagnosed patients (around 60% in the hospital cancer registry of Institut Bordet). The proportions of patients with adenocarcinoma or with squamous cell histology are also heterogeneous from one study to another; one study has been fully dedicated to patients with pure or mixed bronchioloalveolar carcinoma. Most of the studies (18/24) used SUV max for analyzing the relationship between tumor metabolic activity and survival as reported in Table 1.

All 24 eligible studies were scored according to the scale, we previously published.[4] This score includes two subscores: one for the clinical part of the study (items related to the report of the description of the patients population, the study design, the description of therapeutic characteristics, and the description of the radiologic assessments) and one for the PET report (looking at the technical parameters for SUV assessment and patients selection). Considering the 24 studies altogether, the median overall score was 56%, ranging from 27% to 68%. Recently published studies did not get significantly higher scores than the older publications (medians of 54% versus 57%, $P = 0.25$). The median clinical score was 60% (ranging from 34% to 80%) without detectable difference between the two series of publications (median of 55% for the newer ones versus 61% for the older ones, $P = 0.39$). Finally, the median value of the score describing the characteristics of the PET scan methodology was 51% (from 5% to 65%) with medians of 48% versus 53% for recent versus older publications, respectively ($P = 0.78$). The clinical, PET, and total scores are presented in Table 2.

Evaluable Studies

We considered as inevaluable and excluded from further analyses three publications. The first one,[29] because we thought that the analysis was potentially biased. Indeed, in that study including surgically treated patients, all patients recurring locally were excluded from the analysis because the

TABLE 2. Methodological Scores for the 24 Eligible Studies

Study	No. of Patients	Clinical Score	PET Score	Total Score
Ahuja et al.[5]	155	71	50	61
Sugawara et al.[6]	38	64	58	61
Vansteenkiste et al.[7]	125	75	35	56
Dhital et al.[8]	77	34	33	33
Higashi et al.[9]	57	50	60	55
Jeong et al.[10]	73	71	65	68
Downey et al.[11]	100	52	38	45
Borst et al.[a][12]	51	61	53	57
Cerfolio et al.[b][13]	315	71	60	66
Port et al.[14]	64	48	5	27
Sasaki et al.[15]	162	59	53	56
Eschmann et al.[16]	137	80	48	64
Prevost et al.[17]	120	61	53	57
Raz et al.[23]	36	52	25	44
de Jong et al.[24]	66	73	48	51
Downey et al.[25]	487	55	5	31
Lee et al.[26]	19	55	58	56
Na et al.[27]	57	68	63	66
van Baardwijk et al.[28]	102/46[c]	34	53	43
Vesselle et al.[29]	208	61	60	61
Zhang et al.[30]	82	52	58	54
Goodgame et al.[31]	136	52	13	33
Hanin et al.[32]	96	63	43	56
Hoang et al.[33]	214	52	48	50

[a] Study considered as not fully eligible because of inclusion of 11 patients without histological proof of lung cancer but included in the primary review.[4]

[b] Study considered as not fully eligible in the primary review[4] because it was unclear that histological proof of lung cancer was obtained for all patients (but first author confirmed later that it was the case).

[c] Fifty-six of these 102 patients were included in Ref 7.

The scores quantify the information available in the studies regarding the methodology they used at the clinical level and at the imaging level with [^{18}F]-fluoro-2-deoxy-d-glucose positron emission tomography (FDG-PET).

PET, positron emission tomography.

authors considered, in such a situation, that the surgical treatment was inadequate. However, there might be an association between local recurrence and SUV value with a potential bias in the assessment of the relationship between survival and SUV induced by the selection. As the number of patients excluded from the analysis was not specified, we preferred to classify this study as inevaluable. The two other discarded studies were the study by Na et al.[27] and by Lee et al.[26] because reported data were insufficient to extract a HR estimate (in one case SUV was assessed as a continuous covariate, in the other case, it was only reported that no relationship between SUV and survival was identified). The three studies included together 284 patients (9.7% of the total number of patients registered in the evaluable studies which is 2922). The corresponding authors of Refs. 27 and 26 were contacted but did not reply[27] or could not provide the requested data.[26]

Consideration of Overlaps Between Patient Cohorts

Two publications were reported by Downey et al.[11,25] with overlapping accrual periods. The author was contacted and replied that about 35 patients of 100 from the first study were likely to be included in the second one. As the overlap was not total, we reported our results with and without the first study by Downey et al.

As the cohorts reported by Guo et al.[34] and by Higashi et al.[35] in 2000 were fully included in the patients populations published by Higashi et al.[9] in 2002, we even did not consider as eligible the publications by Guo and Higashi (2000).

van Baardwijk et al.[28] reported a pooled analysis of patients recruited in two different institutions, one cohort from Leuven and one cohort from Maastricht. We got confirmation that the Leuven's cohort is a subgroup of the patients population published earlier by Vansteenkiste et al.,[7] and we excluded the patients from Leuven from the meta-analysis as the results from the two institutions were reported separately in the publication by Van Baardwijk et al.[28]

Ahuja et al.[5] and Hoang et al.[33] reported on two patients cohorts diagnosed during intersecting periods in the same institution. The corresponding author (N. Patz) warranted that there was no or minimal overlap between the two cohorts, and we kept the two studies in our meta-analysis.

Individual Results

Table 3 shows the 21 evaluable studies that were included in the meta-analysis with the threshold of SUV used in each series and the method we applied for retrieving an estimate of the HR between the group with high SUV values and the group with low SUV values. It also reports on the individual conclusions regarding the association between SUV and disease-free survival or overall survival. As it was already the case in our previous review, the thresholds used by the authors were heterogeneous, and this heterogeneity is explained by many factors including: the type of PET machine, the algorithms for iteration and reconstruction, the type of correction (weight, lean body mass, etc.), the time elapsed between FDG injection and emission scan, the type of uptake use (mean, max, etc.), the patients populations, and the method for threshold determination (median, the so called best cut-off method, etc.).

Most of the studies identified high values of SUV as being of poor prognosis. Only four studies, three relatively small sized and one dedicated to advanced stages failed to identify a prognostic impact of SUV.

Table 4 reports the individual HR estimates together with their CIs at 95%. By convention, a HR higher than 1 means a worse prognosis for patients with high SUV values. Figure 1 shows the graphical representation of the association between effect size and study size. In the absence of publication bias, this plot should look like a funnel, which seems to be the case for the studies we identified for our review. Indeed, it is expected that the HRs from individual studies should converge toward the true HR when sample size is increasing. Therefore, we should observe a larger spread for small sample sizes than for larger sample sizes with a symmetric pattern. Figure 2 provides a graphical representation of the point estimates of the HR from the individual studies together with their CI at the 95% level.

TABLE 3. Association Between SUV and DFS/OS: Results of Individual Studies

Author	Threshold	Type of Threshold	N	N Low SUV	N High SUV	Rate Low SUV	Extraction	DFS	OS
Ahuja	10.0	Best cut-off	155	118	37	76	Logrank	NA	Poor
Sugawara	8.7	Median	38	19	19	50	Logrank	NA	Unconclusive
Vansteenkiste	7.0	Best cut-off	125	31	94	25	Logrank	NA	Poor
Dhital	10.0	Arbitrary	77	44	33	57	HR	NA	Unconclusive
Higashi	5.0	Best cut-off	57	41	12	72	Logrank	Poor	Poor
Jeong	7.0	Best cut-off	73	33	40	45	Logrank	NA	Poor
Downey	9.0	Median	100	50	50	50	Curves	NA	Poor
Borst	15.0	Median	51	25	26	49	Logrank	NA	Poor
Cerfolio	10.0	Best cut-off	315	162	153	51	Logrank	NA	Poor
Port	2.5	?	64	35	29	55	Logrank	NA	Unconclusive
Sasaki	5.0	Best cut-off	162	43	119	27	Curves	Poor	Poor
Eschmann	12.0	Best cut-off	137	99	38	72	Logrank	NA	Poor
Prevost	10.0	Literature validation	120	99	21	83	Logrank	NA	Poor
Raz	2.5	?	36	19	17	53	HR	NA	Poor
de Jong	6.4	Median + best cut-off	66	33	33	50	HR	NA	Poor
Downey	5.3	Median	487	243	244	50	Logrank	NA	Poor
van Baardwijk	11.0	Best cut-off	46	37	33	80	Logrank	NA	Poor
Zhang	11.0	Median	82	41	41	50	Logrank	Poor	Poor
Goodgame	5.5	Median	136	65	71	48	Logrank	Poor	Poor
Hanin	7.8	Median	97	48	49	49	Logrank	Poor	Poor
Hoang	11.1	Median	214	106	108	50	Logrank	NA	Unconclusive

SUV, standardized uptake value; DFS, disease-free survival; OS, overall survival; HR, hazard ratio; NA, not available.

TABLE 4. Individual HR Estimates for the 21 Evaluable Studies

Author	N	HR	LB 95% CI	UB 95% CI
Ahuja	155	2.05	1.24	3.37
Sugawara	38	0.56	0.21	1.44
Vansteenkiste	125	2.72	1.50	4.94
Dhital	77	1.30	0.70	2.60
Higashi	57	6.20	1.34	28.75
Jeong	73	4.33	1.80	10.45
Downey	100	2.60	1.02	6.64
Borst	51	3.15	1.59	6.22
Cerfolio	315	2.65	1.63	4.31
Port	64	2.36	0.24	22.88
Sasaki	162	7.66	1.41	41.50
Eschmann	137	1.71	1.00	2.93
Prévost	120	2.36	1.34	4.15
Raz	36	9.90	1.20	79.40
De Jong	66	2.93	1.21	7.09
Downey	487	1.58	1.05	2.40
van Baardwijk		3.40	1.40	8.26
Zhang	82	2.36	1.37	4.06
Goodgame	136	1.89	1.20	2.99
Hanin	96	2.83	1.52	5.26
Hoang	214	1.29	0.94	1.76

HR, hazard ratio; LB, lower bound; UB, upper bound; CI, confidence interval.

Survival Data Aggregation

Considering all together the 21 studies ($n = 2637$), we detected presence of heterogeneity (χ^2 statistic = 36.11, 20 *df*, $P = 0.01$); this heterogeneity can be reduced by discarding the publication by Sugawara et al.[6] with a P value becoming 0.06. Using a random-effects model, we reached a combined HR of 2.08, significantly different from 1 with a 95% CI ranging from 1.69 to 2.56. If we restrict the analysis to the newer studies, we obtained, with a fixed effects model, a combined HR of 1.80 (95% CI, 1.50–2.16) and using a random effects model, a pooled estimate of 2.03 (95% CI, 1.53–2.70).

When dividing the studies, according to their publication date, we obtained a combined HR of 2.03 (95% CI, 1.53–2.70, random-effects model) for the newer studies (identified after the publication of our first review). There was no interaction between the older and the newer publications ($P = 0.60$).

Excluding the study having included some patients without histologic proof of lung cancer[12] and the first study by Downey et al.[11] with overlapping patients, the combined HR became 2.11 (95% CI, 1.71–2.59) with presence of heterogeneity ($P = 0.01$).

Excluding the 7 studies having determined the threshold for the SUV value using the wrongly called best cut-off method, the combined HR decreased to 1.97 (95% CI, 1.53–2.52) with a P value for the test for heterogeneity of 0.02.

We further focused on studies having included only nonmetastatic patients. Fourteen studies were in this situation. Heterogeneity was not detectable any more ($P = 0.26$). Combined HRs for fixed-effects and random-effects

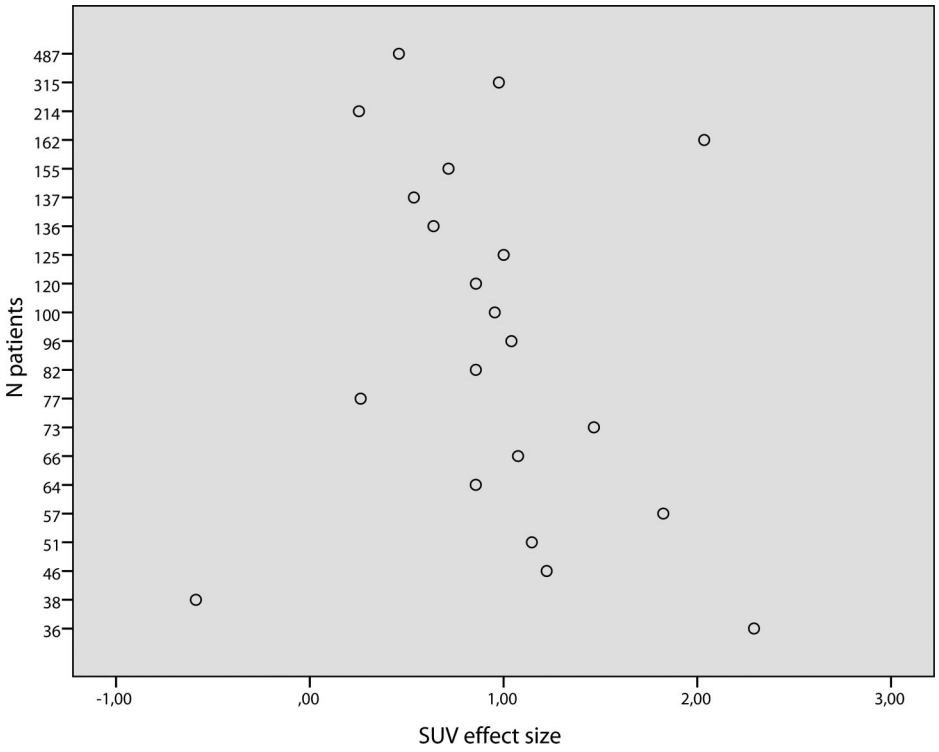

FIGURE 1. Funnel plot. Effect size: neperian logarithm of the hazard ratio.

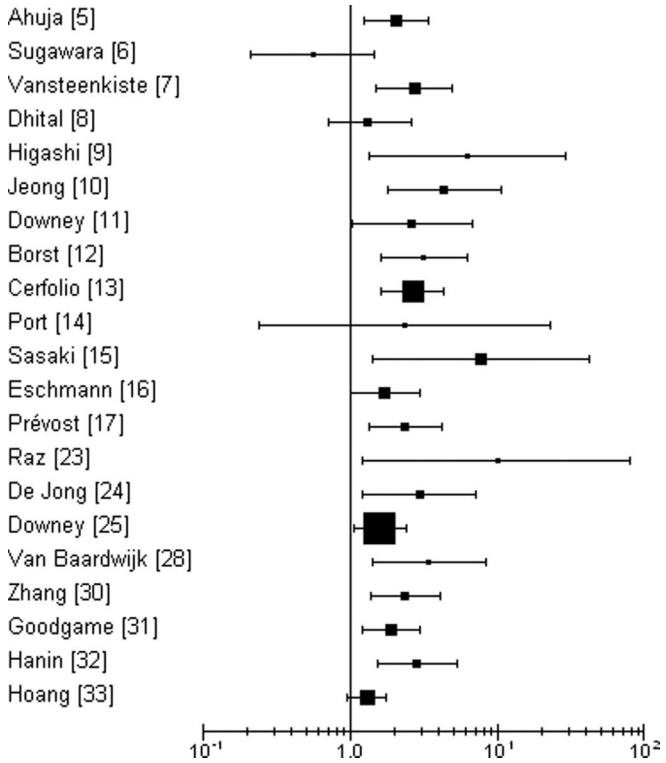

FIGURE 2. Individual hazard ratios with their 95% confidence interval. Point estimates of individual hazard ratios with their 95% confidence interval on a logarithmic scale. By convention, HR >1 means that patients with a higher standard uptake value (SUV) on the primary tumor have a worse prognosis.

models were 2.18 (95% CI, 1.83–2.60) and 2.25 (95% CI, 1.84–2.75), respectively. Without considering the studies with best cut-off thresholds, 10 cohorts are remaining in the analysis without detectable heterogeneity ($P = 0.25$) and combined HR of 2.11 (95% CI, 1.72–2.59) and 2.18 (95% CI, 1.72–2.77).

It is noteworthy that only one study was dedicated to advanced stages with 214 patients and 158 observed events that failed to detect a HR significantly different from 1. In the subgroup of studies with unselected series for disease stage, the combined HRs for a fixed-effects model and a random-effects model were 1.76 (95% CI, 1.44–2.15) and 1.90 (95% CI, 1.27–2.84), respectively.

There was, however, no significant interaction according to the fact that the authors included or excluded patients with stage IV disease ($P = 0.46$).

As intended in our plan for analysis, we looked for subgroups according to histology, but we did not find any single study reporting separately on patients with adenocarcinoma or squamous cell histology and none of them analyzed the possibility of an interaction between histology and SUV prognostic impact. Few studies looked at some subgroups defined by stage or type of treatment (curative resection for instance) but we did not find consistent subgroups to analyze separately.

Figure 3 reports on a graphical way the different combined HRs that we estimated.

Finally, four studies report on disease free survival, but we believed that these studies were not representative enough of the complete set of studies to justify the calculation of a combined HR for these four studies that individually all identify high SUV value as a poor prognostic factor for disease free survival.

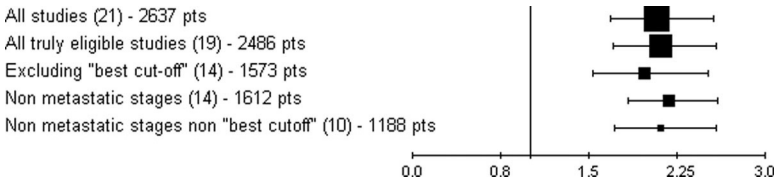

FIGURE 3. Combined HRs for survival results aggregation using different subsets of studies. Combined hazard ratios with 95% confidence intervals obtained from random-effects models.

DISCUSSION

We confirmed, with 10 further studies and about 1000 additional patients, the results of our first review concluding that patients with a tumor demonstrating a higher metabolic activity as measured by SUV have shorter survival than patients with a tumor with a lower glucose metabolic rate. There are only little changes in the point estimates but our narrower CIs give more reliability to the results, especially when we are focusing on studies looking at more restricted eligibility criteria. Of course, all the limitations that we addressed in our previous report and in our other systematic reviews looking at prognostic factors for survival in lung cancer are still present. We were not able to assess the independent prognostic value of SUV, and this question remains open with contradictory results in the studies that we identified: for some of them, the prognostic value is kept after adjustment for other prognostic features and for some of them, the SUV is left out of adjusted models (for instance in Downey et al.[25] once pathologic staging is taken into account). The methodology for assessing SUV and the type of FDG-PET machine are obviously extremely important as there are various measurement errors that can occur, and we already insisted on that aspect (we refer the interested reader to appendixes 1 and 2 of our previous review[4]). There are currently some researchers starting to work on standardization of assessment methods (conditions of the examination, calibration of the machines, choice of reconstruction and attenuation algorithms, etc.) and this standardization is of course crucial for the conduct of multicentric studies that are most likely needed to establish the independent prognostic value of SUV. although it does not solve completely the issue of reproducibility of results.[36]

Publication bias is another very important issue, but we did not detect any strong bias, although we did not perform a formal test as these tests are known to lack from power.[18] Heterogeneity was detected and should be worth to explain. We identified one study inducing heterogeneity[6] but this study is certainly not only the factor responsible for heterogeneity. The *P* value of the test for heterogeneity increased but remained significant after exclusion of that study. The case mix of patients and the measurement of SUV itself are other factors certainly contributing to the heterogeneity but with the data available in the publications, we were unable to define subgroups of publications more homogeneous and this justifies our use of random-effects models. Our results, however, suggest that the prognostic impact of SUV might be stage dependent, and it is indeed possible that once a tumor has advanced characteristics, the anatomic extent of the tumor is so important for predicting the prognosis that the metabolic activity measured by SUV max on the primary tumor has lower prognostic value. This is why the introduction of other measures of metabolic activity like the tumor lesion glycolysis might be very interesting.

Our results do not allow to conclude to an optimal threshold but only that higher values of SUV imply higher hazards and some authors do in fact suggest that there are not two groups of patients but rather that there is a continuous increase in the hazard as SUV increases.[24,37] We also know that SUV is linked to histology and different behaviors may apply for patients with adenocarcinoma and for patients with squamous cell lung cancer. Unfortunately, we were not able to investigate that question as we had to deal with published results. However, some authors tried to find a "best" discriminant threshold but without correcting for multiple analyses,[38] and it has been well described in the literature that such methods may lead to the identification of false positive prognostic factors.[19] As far as we could, we used the median value of the SUV distribution in each individual study and we addressed the potential bias by considering, as a separate subgroup, studies, which used a method with better reproducibility chances. The combined HR decreased, as expected, but remained significantly different from one and represents likely a less biased estimate of the true impact of SUV.

Some of the limitations of the review that we performed could be addressed if we were able to conduct a meta-analysis based on individual patients data. In that situation, it would be possible to explore the associations between SUV and other covariates of interest, mainly stage, and histology and, of outstanding importance, to assess the independent prognostic value of SUV. It would be worth also to look, in a per study stratified analysis, at SUV as a continuous covariate although this raises, again, the issue of standardization of PET assessment. Therefore, we plan to compare the different PET algorithms used by the study coordinators to define, if possible, groups of studies with more homogeneous methodology for the measure of the metabolic activity. It would also be possible to better select the patients by excluding overlapping information and restricting the analysis to patients with histologic proof of cancer. Finally, updating the survival data would also be beneficial. After our first review, we contacted all the corresponding authors for proposing such a meta-analysis, but despite the interest shown by many of them for such a project, we did not yet collect a sufficient number of data bases to conduct an individual patients' data. The project will now be endorsed by the IASLC staging project and we intent, in the following weeks, to contact again authors of the "older" publications and to get in touch the authors of the more recent publications on behalf of the IASLC. We believe that such a cooperation and such an analysis would allow to obtain new and interesting results and, ultimately, to plan a prospective prognostic factors study with an a priori calculated sample size, which would be the last step to provide definite results about the interest to integrate a measure of the metabolic activity of the tumor in prognostic models.

ACKNOWLEDGMENTS

We thank the authors who provided to us the way to estimate the individual HR from their studies, A. Van Baardwijk, J. Van Steenkiste; those who gave us further information about the potential overlaps between patients cohorts, B. Downey, N. Patz, and J. Van Steenkiste. We also thank R. Cerfolio who clarified the patient selection for the cohort he considered.

REFERENCES

1. Paesmans M. [Prognostic factors in lung cancer.] *Rev Mal Respir* 2005;22:8S76–8S80 [in French].
2. Sculier JP, Chansky K, Crowley JJ, et al. The impact of additional prognostic factors on survival and their relationship with the anatomical extent of disease expressed by the 6th Edition of the TNM Classification of Malignant Tumors and the proposals for the 7th Edition. *J Thorac Oncol* 2008;3:457–466.
3. Gould MK, Kuschner WG, Rydzak CE, et al. Test performance of positron emission tomography and computed tomography for mediastinal staging in patients with non-small-cell lung cancer: a meta-analysis. *Ann Intern Med* 2003;139:879–892.
4. Berghmans T, Dusart M, Paesmans M, et al. Primary tumor standardized uptake value (SUVmax) measured on fluorodeoxyglucose positron emission tomography (FDG-PET) is of prognostic value for survival in non-small cell lung cancer (NSCLC): a systematic review and meta-analysis (MA) by the European Lung Cancer Working Party for the IASLC Lung Cancer Staging Project. *J Thorac Oncol* 2008;3:6–12.
5. Ahuja V, Coleman RE, Herndon J, et al. The prognostic significance of fluorodeoxyglucose positron emission tomography imaging for patients with nonsmall cell lung carcinoma. *Cancer* 1998;83:918–924.
6. Sugawara Y, Quint LE, Iannettoni MD, et al. Does the FDG uptake of primary non-small cell lung cancer predict prognosis? A work in progress. *Clin Positron Imaging.* 1999;2:111–118.
7. Vansteenkiste JF, Stroobants SG, Dupont PJ, et al. Prognostic importance of the standardized uptake value on (18)F-fluoro-2-deoxy-glucose-positron emission tomography scan in non-small-cell lung cancer: an analysis of 125 cases. Leuven Lung Cancer Group. *J Clin Oncol* 1999;17:3201–3206.
8. Dhital K, Saunders CA, Seed PT, et al. [(18)F]Fluorodeoxyglucose positron emission tomography and its prognostic value in lung cancer. *Eur J Cardiothorac Surg* 2000;18:425–428.
9. Higashi K, Ueda Y, Arisaka Y, et al. 18F-FDG uptake as a biologic prognostic factor for recurrence in patients with surgically resected non-small cell lung cancer. *J Nucl Med* 2002;43:39–45.
10. Jeong HJ, Min JJ, Park JM, et al. Determination of the prognostic value of [(18)F]fluorodeoxyglucose uptake by using positron emission tomography in patients with non-small cell lung cancer. *Nucl Med Commun.* 2002;23:865–870.
11. Downey RJ, Akhurst T, Gonen M, et al. Preoperative F-18 fluorodeoxyglucose-positron emission tomography maximal standardized uptake value predicts survival after lung cancer resection. *J Clin Oncol* 2004; 22:3255–3260.
12. Borst GR, Belderbos JS, Boellaard R, et al. Standardised FDG uptake: a prognostic factor for inoperable non-small cell lung cancer. *Eur J Cancer* 2005;41:1533–1541.
13. Cerfolio RJ, Bryant AS, Ohja B, et al. The maximum standardized uptake values on positron emission tomography of a non-small cell lung cancer predict stage, recurrence, and survival. *J Thorac Cardiovasc Surg* 2005;130:151–159.
14. Port JL, Andrade RS, Levin MA, et al. Positron emission tomographic scanning in the diagnosis and staging of non-small cell lung cancer 2 cm in size or less. *J Thorac Cardiovasc Surg* 2005;130:1611–1615.
15. Sasaki R, Komaki R, Macapinlac H, et al. [18F]fluorodeoxyglucose uptake by positron emission tomography predicts outcome of non-small-cell lung cancer. *J Clin Oncol* 2005;23:1136–1143.
16. Eschmann SM, Friedel G, Paulsen F, et al. Is standardised (18)F-FDG uptake value an outcome predictor in patients with stage III non-small cell lung cancer? *Eur J Nucl Med Mol Imaging* 2006;33:263–269.
17. Prevost S, Boucher L, Larivee P, et al. Bone marrow hypermetabolism on 18F-FDG PET as a survival prognostic factor in non-small cell lung cancer. *J Nucl Med* 2006;47:559–565.
18. Egger M, Davey SG, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629–634.
19. Altman DG, Lausen B, Sauerbrei W, et al. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst* 1994;86:829–835.
20. Na II, Byun BH, Kang HJ, et al. 18F-fluoro-2-deoxy-glucose uptake predicts clinical outcome in patients with gefitinib-treated non-small cell lung cancer. *Clin Cancer Res* 2008;14:2036–2041.
21. Davies A, Tan C, Paschalides C, et al. FDG-PET maximum standardised uptake value is associated with variation in survival: analysis of 498 lung cancer patients. *Lung Cancer* 2007;55:75–78.
22. Lee KH, Lee SH, Kim DW, et al. High fluorodeoxyglucose uptake on positron emission tomography in patients with advanced non-small cell lung cancer on platinum-based combination chemotherapy. *Clin Cancer Res* 2006;12:4232–4236.
23. Raz DJ, Odisho AY, Franc BL, et al. Tumor fluoro-2-deoxy-D-glucose avidity on positron emission tomographic scan predicts mortality in patients with early-stage pure and mixed bronchioloalveolar carcinoma. *J Thorac Cardiovasc Surg* 2006;132:1189–1195.
24. de Jong WK, van der Heijden HF, Pruim J, et al. Prognostic value of different metabolic measurements with fluorine-18 fluorodeoxyglucose positron emission tomography in resectable non-small cell lung cancer: a two-center study. *J Thorac Oncol* 2007;2:1007–1012.
25. Downey RJ, Akhurst T, Gonen M, et al. Fluorine-18 fluorodeoxyglucose positron emission tomographic maximal standardized uptake value predicts survival independent of clinical but not pathologic TNM staging of resected non-small cell lung cancer. *J Thorac Cardiovasc Surg* 2007; 133:1419–1427.
26. Lee P, Weerasuriya DK, Lavori PW, et al. Metabolic tumor burden predicts for disease progression and death in lung cancer. *Int J Radiat Oncol Biol Phys* 2007;69:328–333.
27. Na II, Cheon GJ, Choe DH, et al. Clinical significance of (18)F-FDG uptake by N2 lymph nodes in patients with resected stage IIIA N2 non-small-cell lung cancer: a retrospective study. *Lung Cancer* 2008; 60:69–74.
28. van Baardwijk A, Dooms C, van Suylen RJ, et al. The maximum uptake of (18)F-deoxyglucose on positron emission tomography scan correlates with survival, hypoxia inducible factor-1alpha and GLUT-1 in non-small cell lung cancer. *Eur J Cancer* 2007;43:1392–1398.
29. Vesselle H, Freeman JD, Wiens L, et al. Fluorodeoxyglucose uptake of primary non-small cell lung cancer at positron emission tomography: new contrary data on prognostic role. *Clin Cancer Res* 2007;13:3255–3263.
30. Zhang ZJ, Chen JH, Meng L, et al. 18F-FDG uptake as a biologic factor predicting outcome in patients with resected non-small-cell lung cancer. *Chin Med J* (*Engl*) 2007;120:125–131.
31. Goodgame B, Pillot GA, Yang Z, et al. Prognostic value of preoperative positron emission tomography in resected stage I non-small cell lung cancer. *J Thorac Oncol* 2008;3:130–134.
32. Hanin FX, Lonneux M, Cornet J, et al. Prognostic value of FDG uptake in early stage non-small cell lung cancer. *Eur J Cardiothorac Surg* 2008;33:819–823.
33. Hoang JK, Hoagland LF, Coleman RE, et al. Prognostic value of fluorine-18 fluorodeoxyglucose positron emission tomography imaging in patients with advanced-stage non-small-cell lung carcinoma. *J Clin Oncol* 2008;26:1459–1464.
34. Guo J, Higashi K, Yokota H, et al. In vitro proton magnetic resonance spectroscopic lactate and choline measurements, 18F-FDG uptake, and prognosis in patients with lung adenocarcinoma. *J Nucl Med* 2004;45: 1334–1339.
35. Higashi K, Ueda Y, Ayabe K, et al. FDG PET in the evaluation of the aggressiveness of pulmonary adenocarcinoma: correlation with histopathological features. *Nucl Med Commun* 2000;21:707–714.
36. Boellaard R, Oyen WJG, Hoekstra CJ, et al. The Netherlands protocol for standardisation and quantification of FDG whole body PET studies in multi-centre trials. *Eur J Nucl Med Mol Imaging* 2008;35:2320–2333.
37. Detterbeck FC, Vansteenkiste JF, Morris DE, et al. Seeking a home for a PET, Part 3. Emerging applications of positron emission tomography imaging in the management of patients with lung cancer. *Chest* 2004;126: 1656–1666.
38. Hilsenbeck SG, Clark GM. Practical p-value adjustment for optimally selected cutpoints. *Stat Med* 1996;15:103–112.

