The role of FDG-PET in Thyroid Nodules with Indeterminate Fine-Needle Aspiration Biopsy: systematic review and meta-analysis of the literature

D. Vriens, MD*; J.H.W. de Wilt, MD, PhD#; G.J. van der Wilt, MD, PhD†; R.T. Netea-Maier, MD, PhD¶; W.J.G. Oyen, MD, PhD* and L.F. de Geus-Oei, MD, PhD*.

Departments of *Nuclear Medicine, #Surgery, †Epidemiology, Biostatics and Health Technology Assessment and ¶Endocrinology, Radboud University Nijmegen Medical Centre

Running Head: FDG-PET in indeterminate thyroid nodules.

Correspondence:

D. Vriens, MD

Department of Nuclear Medicine (internal postal code 444)

Radboud University Nijmegen Medical Centre

P.O. Box 9101, 6500 HB Nijmegen, the Netherlands

Phone: +31-24-3614048, Fax: +31-24-3618942

E-mail: D.Vriens@nucmed.umcn.nl

This document contains 27 pages, 3 tables and 5 figures.

Word-count all text: 6146

There are no financial disclosures for any of the authors.

Condensed abstract: In this meta-analysis of 6 studies including 225 patients with thyroid nodules with indeterminate FNAB, FDG-PET proved an excellent discriminator for which patients do not benefit from diagnostic surgery as the negative predictive value to rule out thyroid cancer was 96%.

Abstract

Background: Indeterminate results at fine needle aspiration biopsy (FNAB) of thyroid nodules pose a clinical dilemma as only 20-30% of the patients suffer from malignancy. Previous studies suggested that the false-negative ratio of FDG-PET is very low, therefore it might help to identify those who benefit from (hemi)thyroidectomy.

Methods: A systematic search was performed in five databases. After assessment, the studies were analyzed for heterogeneity and the extracted data of test characteristics were pooled using a random effects model. Threshold effects were examined and publication bias was assessed.

Results: The query resulted in 239 records, of which 6 studies met our predefined inclusion criteria. The data of 225 of the 241 described patients could be extracted. There was mild to moderate heterogeneity in study results (I^2=0.390-0.867). The pooled prevalence of malignancy was 26%. Pooled sensitivity, specificity, positive and negative predictive value and accuracy were 95% (95%-confidence interval: 86-99%), 48% (40-56%), 39% (31-47%), 96% (90-99%) and 60% (53-67%), respectively. Sensitivity increased to 100% for the 164 lesions larger than 15mm. There were no evidences of threshold effects or publication bias. **Conclusion:** A negative FDG-PET scan in thyroid nodules larger than 15 mm with indeterminate FNAB excludes thyroid cancer in a pooled population of 225 patients. Conversely, a positive FDG-PET does not implicate cancer as approximately half of these patients have benign nodules. Therefore, incorporation of FDG-PET in the initial work-up of these patients prior to surgery is worth further investigation.

Keywords: Fluorodeoxyglucose F18, Positron-Emission Tomography, Meta-Analysis, Systematic Review, Thyroid Nodules, Thyroidectomy.

Introduction

Thyroid cancer is the most common endocrine malignancy. It represents approximately 1% of all cancers, corresponding to an incidence of up to 10.2 per 100 000 people per year in the US.[1, 2] with increasing incidence over the last decades.[2, 3]

The most common clinical presentation of thyroid cancer is a thyroid nodule (TN), either solitary or (dominant) within a multinodular goiter. Approximately 5-10% of adults have palpable TNs and 17-45% have nodules identified by ultrasound.[4-6] The majority of these nodules are benign, but approximately 5-15% of all palpable thyroid nodules are malignant.[5, 7] Fine-needle aspiration biopsy (FNAB) is the most important diagnostic test in the initial evaluation of a patient with a TN with a high diagnostic accuracy (70-97% in experienced centers). Approximately 70% of the results of FNAB are classified as benign, 4% as malignant, 2-10% as insufficient material and the remaining as either indeterminate or suspicious (16-24%). Due to similar cytological features, it is particularly challenging to distinguish between different types of thyroid neoplasms of the follicular type (i.e. follicular thyroid adenoma, follicular thyroid carcinoma (FTC) and follicular type of papillary carcinoma). Therefore, patients with indeterminate or suspicious FNAB results have to undergo diagnostic hemithyroidectomy to exclude malignancy.[7] As only 20-30% of these nodules are malignant[8], most of the patients are unnecessarily undergoing thyroid surgery, with potential risk for irreversible complications.

Several promising markers (e.g. thyroid peroxidase (TPO), galectin-3, telomerase, RET/PTC, p53) have been studied in patients with TNs in order to improve the accuracy of FNAB. Yet none of these markers have reached routine clinical use as they have only been documented in a subset of tumors.[9] Both ultrasonographic and scintigraphic features of TNs have been investigated in the past for their diagnostic value in the preoperative diagnostic work-up of patients with thyroid nodules, but none of these technique could accurately distinguish between benign and malignant nodules.

Characterization of tissue using the glucose analogue [^{18}F]-2-fluoro-2-deoxy-D-glucose (FDG) together with positron emission tomography (PET) has proven beneficial in diagnostics and follow-up of many malignancies.[10] The use of FDG-PET/CT in the management of thyroid disease has been

limited primarily to the postoperative surveillance of patients with known differentiated thyroid carcinoma (postoperative staging for remnant disease and in therapy response assessment[5]). There is a special role for FDG-PET/CT in postoperative surveillance in case thyroglobulin (Tg) is elevated but whole-body [^{131}I]$^{-}$scintigraphy is negative.[5] Finally, FDG-PET/CT thyroid incidentalomas are found in approximately 1-2% of FDG-PET/CTs and harbor a 14-47% chance of being confirmed malignant[11-13] warranting further investigations.[5, 14] Currently, there is no routine place for FDG-PET/CT in the work-up of a TN.[5]

This systematic review aims to provide an up-to-date summary of the value of FDG-PET/CT for the preoperative evaluation of patients with TNs and either indeterminate or repeatedly insufficient FNAB. By a systematic literature search and meta-analysis, the false-negative rate of FDG-PET/CT is quantified to investigate whether a negative FDG-PET/CT scan can select patients with a low suspicion of malignancy in whom surgery therefore can be omitted.

Material and Methods

Literature Search and Study Selection

A systematic search was performed using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement (Figure 1).[15, 16]

The main research question was defined using the PICTS strategy: target Population (including previous tests), Index test, Comparator test, Target condition and Study design.[17-19] It was formulated into a search query containing a combination of Medical Subject Headings (MeSH) or keywords and truncated synonyms (Boolean operators). Then, a search using this query was performed November 21 2010, with the following five search engines, which use partially overlapping databases: PubMed, Scopus, the library of the Cochrane Collaborations, OvidSP MEDLINE and OvidSP MEDLINE In Process & Other Non-Indexed Citations. Review articles and letters-to-the editor, articles with less than 5 patients included and articles written in languages other than English, French, German or Dutch were excluded.

Quality Assessment and Data Extraction

Quality appraisal of retrieved full-text articles were all graded independently by two investigators, for quality and applicability by the Quality Assessment tool for Diagnostic Accuracy Studies (QUADAS). This widely used tool consists of 14 items that cover patient spectrum, reference standard, disease progression bias, verification and review bias, clinical review bias, incorporation bias, test execution, study withdrawals and intermediate results.[20, 21] Disagreements were resolved by consensus after re-evaluation of the references.

Individual patient data were extracted from the approved published articles for all patients with indeterminate (follicular or Hürthle cell (oxyphilic, oncocytic) proliferation) or (repeatedly) inconclusive FNAB in whom both an FDG-PET and surgery was performed and final histopathological diagnosis was available. Heterogeneity in PET acquisition and quantification of

FDG-uptake led to the conclusion that between study comparison of quantitative parameters was impossible[22] and therefore not attempted. Therefore, a positive test result was solely based on visual assessment of the PET-scan and defined as "any focal increased uptake in the region of the TN above background". Final histopathological diagnosis served as the Gold Standard outcome parameter. When final histopathology showed microcarcinoma (defined as <1 cm in diameter) this was considered a coincidental finding in benign disease. Maximum lesion diameter measured on histopathology was noted for all patients. As the diameter of the lesions previously described by our group were not published[23], they were retrieved from the original data. In case of multiple nodules, bias caused by FNAB sampling error or spatial mismatch between the nodule and on the FDG-PET was neglected. Therefore a patient-based rather than a lesion-by-lesion analysis was performed.

Data Synthesis and Statistical Analysis:

Agreement in the per-item QUADAS score per reference between the two reviewers was expressed using Cohen's κ-coefficient and was interpreted according to the suggestions of Landis and Koch.[24] The intraclass correlation coefficient (ICC) using a two-way random effects model with a definition of absolute agreement was used to quantify agreement for the overall score.

Heterogeneity of the study populations was assessed by comparing the distribution of possible confounders in the included references. Proportions (sex, FNAB results, prevalence of malignancy, histology results) were compared between references using either the χ^2-statistic or Fisher's exact test. The continuous potential confounders (patient age, tumor diameter) were assessed for (log-)normality (histograms, skewness, kurtosis, Shapiro-Wilk). In case of (log-)normality means (±standard deviations, SD) and in all other cases medians (with interquartile ranges, IQR) are presented. Comparison of (log-)normal distributed variables between references was performed by one-way ANOVA followed by Tukeys HSD post-hoc test. In case of violation of normality, Kruskal-Wallis H was used as non-parametric equivalent for comparison between multiple independent groups and the Mann-Whitney U test for comparison between two independent groups.

Extracted data was ordered in 2x2 contingency tables from which disease prevalence and diagnostic test characteristics (Sensitivity, Specificity, Positive (PPV) and Negative Predictive Value (NPV), Positive (LR+) and Negative Likelihood Ratio (LR-), diagnostic Odds Ratio (dOR) and Accuracy (Acc)) could be calculated using the classical equations. To avoid calculation problems by having zero values, 0.5 was added to each cell of the respective contingency table as is commonly used.[25] 95-confidence intervals (95-CI) for the proportions were calculated using the β-distribution (the exact Clopper-Pearson interval), since the commonly used asymptotic normal approximation only holds true for observed frequencies higher than five individual patients.[26]

Since sensitivity and specificity are often inversely related because of the threshold effect, study heterogeneity in these diagnostic test characteristics was visualized using a summary Receiver Operating Characteristic (sROC) curve for which the Area Under the Curve (AUC) was calculated.[27, 28] Other causes of between-study heterogeneity in diagnostic test characteristics were assessed using χ^2 statistics (heterogeneity was defined as $p<0.10$) and quantified by the inconsistency index (I^2, i.e. the amount of variability in the results attributable to between-study variation[29]). A funnel plot of the dOR of each study is constructed to provide insight into publication bias.[30] Since the dOR is approximately log-normally distributed, correlation between the log(dOR) and the standard error of the study effect determined by the Kendall τ_b rank correlation coefficient[31] and the Egger statistic[32], as a significant correlation suggests publication bias.

Pooled sensitivity and specificity were computed based on individual patient data (i.e. the using the sum of the true positive, true negative, false positive and false negative individuals). Ratios (LR+, LR- and dOR) were pooled using 3 strategies: (1) pooling individual patient data, which assumes negligible heterogeneity, (2) weighted (Mantel-Haenzel) averaging of per-study data, which assumes a fixed effect and weights studies by their precision and therefore is less sensitive to (small) inaccurate studies with extreme effects and (3) weighted (DerSimonian-Laird) averaging of per-study data, which allows a random effects and therefore is less sensitive to study heterogeneity. Data is presented in forest plots and τ^2 is presented as a value for the between-study variance of the random effects model.

All analyses were performed using SPSS version 16.0.2. A two tailed $p<0.05$ was considered significant. All meta-analyses (pooling and sROC analysis) are performed using Meta-DiSc version 1.4.[33]

Results

Literature Search and Study Selection

After removal of duplicates, the query resulted in 239 articles of which, after discarding references fulfilling the exclusion criteria, 29 references remained (Figure 1). Of these, 20 dealt with "prevalence of malignancy in thyroid incidentalomas found on FDG-PET/CT" and 9 with the main search question.[23, 34-41]

Quality assessment

There was substantial agreement between both reviewers concurring in 104 of 126 QUADAS items (κ=0.693, $p<0.001$). The correlation of absolute overall scores between both reviewers was moderate (ICC=0.721, p=0.010). Three references were unsuitable for meta-analysis due to limitation to lesion-by-lesion analysis[35], lack of definition of FDG-PET-positivity and of final histopathological diagnosis in each patient[39] or being unclear with respect to which and how many patients had indeterminate FNAB results.[40] In none of the reviewed articles the FDG-PET and histology data were interpreted in combination with other clinical data that would be available in practice (QUADAS item 12). Only one reference[35] described the methodology of the FDG-PET in sufficient detail to permit replication (QUADAS item 9) and only one[41] mentioned blinding of the pathologist to the FDG-PET findings (QUADAS item 11).

Study Heterogeneity

The selected studies were carried out in Austria (in an endemic goiter area)[34], the Netherlands[23], South-Korea[36], Brazil[37] and the USA[38, 41]. Four of the studies were carried out in university hospitals[36-38, 41], one in a general hospital[34] and one was a multicentre trial carried out in a university hospital and a general hospital[23]. In all but one study[38], all patients had TSH within the normal range. The portion of patients with a single thyroid nodule varied from 50% to 100%. Population, in- and exclusion

criteria of these 6 studies are described in Table 1. Inclusion criteria varied somewhat among studies, mainly concerning nodule size, preselection of patients with nodules with suspect ultrasonographic or scintigraphic features.[34]

In two studies not all patients could be used for this meta-analysis: 6/43 patient in one study[34] showed papillary thyroid carcinoma (PTC) at FNAB and were reported as positive controls, 10/46 patients in another study[36] refused surgery, therefore no histopathological diagnosis was available. As a result, the data of 225 of 241 individual patients were available for pooling. Results of US investigation was described for 35% of these.

Data for age, sex, FNAB results, FDG-PET results and final histopathological diagnosis was known on the individual patient level. Patient age, maximum nodule diameter measured histologically and FNAB results, all of which are potential confounders causing heterogeneity in study results, were significantly different between the six included studies. The fraction of female patients was not significantly different (Table 2).

FDG-PET was performed in all patients at least 17 days post-FNAB[23, 36, 37], although 3 studies did not mention this interval.[34, 38, 41] It was performed 60-70 min after injection of 188-555MBq of FDG and visually interpreted by one or two (blinded) experienced observers. A positive FDG-PET was defined as any FDG-uptake higher than the background thyroid bed in all but one study[38]. This study added a further restriction that the maximum standardized uptake value (SUV_{max}) of the lesion had to be higher than 2.0 to be considered FDG-PET positive.[38] This quantitative restriction was not considered in the meta-analysis for the purpose of homogenization of definition of a positive test. Therefore patients #5, #13 and #15 in the study of Hales et al.[38] were shifted to PET-positive by our definition. Another study[41] described 4 patients with "incidental" PTC (0.3-17mm in diameter). Due to the fact the lesions were found distant from the nodule of interest as seen on US and since the nodule of interest was caused by another (benign) thyroid disease the nodule was considered benign.

The prevalence of malignancy in the pooled population was 25.8% (range 13.6% to 41.7%, Fishers Exact=10.7, p=0.055) (Table 3). Of the benign disorders, multinodular goiter was the cause of thyroid nodules in 44% (range 15-71%), follicular adenoma in 33% (range 0-52%), Hürthle cell adenoma in 9% (range 0-37%), (lymphocytic) thyreoiditis in 4% (range 0-33%) and a combination of

benign disorders in 10% (range 0-33%). The distribution of benign causes of thyroid nodules was statistically different between studies (Fishers Exact=70.7, p<0.001). Malignant disorders were mainly caused by FTC (35%, range 0-73%), PTC (26%, range 0-50%), Hürthle cell carcinoma (5%, range 0-13%) and anaplastic carcinoma (3%, range 0-20%). The remaining 31% of malignant cases (range 0-60%) were either a combination of malignant histologies, variants of FTC or PTC or malignancy not otherwise specified by the authors. The distribution of malignant causes of thyroid nodules was statistically different between studies (Fishers Exact=38.2, p<0.001).

Pooled Data

The pooled sensitivity of FDG-PET for the detection of cancer was 94.8% (95-CI: 85.6-98.9%), but there was moderate though non-significant inconsistency among studies (I^2=0.390, $\chi 2_{df=5}$=8.2, p=0.146). The pooled specificity was 47.9% (95-CI: 40.1-55.8%), but there was a high and significant inconsistency between studies (I^2=0.867, $\chi^2_{df=5}$=37.6, p<0.001) mainly because a specificity of 0% was reported in one study[36] in which all patients had positive FDG-PET scans (i.e. there were no negative cases). We could not find a (methodological) cause and therefore did not exclude this study from further analysis. The pooled NPV was 96.4% (95-CI: 89.8-99.2%) and the pooled PPV was 38.7% (95-CI: 30.7-47.3%), with an overall accuracy of FDG-PET for determination of thyroid nodule malignancy of 60.0% (95-CI: 53.3-66.5%) (Table 3).

Based on the pooled analysis of individual patient data, the LR+ was 1.82 (95-CI: 1.56-2.13). Using the fixed effects the LR+ was 1.56 (95-CI: 1.36-1.80, high heterogeneity: $\chi^2_{df=5}$=76.0, p<0.001, I^2=0.934), but considering the heterogeneity random effects modeling seemed more appropriate. With random effects the LR+ was 1.67 (95-CI: 0.98-2.84, τ^2=0.39). The LR- could not be computed for one study[36] since there were no FDG-PET negative cases. Therefore, pooling was based on the remaining 5 references. Using fixed effects the LR- was 0.19 (95-CI: 0.076-0.46, low heterogeneity: $\chi^2_{df=4}$=2.33, p=0.676, I^2<0.001). With random effects the LR- was 0.24 (95-CI: 0.10-0.59, τ^2<0.01) (Figure 2). Therefore the pre-FDG-PET probability of malignancy (prevalence) rose from 25.8% to 38.7% for a positive FDG-PET (PPV) and decreased to 3.6% in case the FDG-PET was negative (1-NPV).

The sROC curve showed no 'shoulder arm' plot suggesting no threshold effect (Figure 3), the AUC was 0.84 (±0.079). The symmetrical funnel plot of the dOR showed no evidence of publication bias (Figure 4), which was confirmed by an insignificant correlation (Kendall τ_b=0.067, p=0.851, Egger's statistic: t=-0.054, p=0.959).

In total three of 58 patients with differentiated thyroid carcinoma showed a false-negative FDG-PET scan. Patient #1 in Hales *et al.*[38] was diagnosed with a 14 mm PTC, which showed diffuse moderate uptake (SUV_{max} 1.5) similar to the thyroid background FDG-uptake. Patient #1 and #12 in Traugott *et al.*[41] both showed a follicular variant PTC of 7 and 9 mm respectively, but no focal increase of FDG-uptake was seen in the thyroid gland. These 3 false-negative lesions were significantly smaller in diameter than other malignant lesions (median diameter 26mm (IQR: 18-50mm), Mann-Whitney U: 15.0, p=0.011) as well as all other lesions (median diameter 21mm (IQR: 15-35mm), Mann-Whitney U: 72.0, p=0.014). Evaluation of the 164 TNs larger than 15 mm therefore led to a sensitivity of FDG-PET in detection of thyroid cancer of 100% (95-CI: 92.5-100%), whereas specificity remained similar (46.6%, 95-CI: 37.4-56.0%) (Figure 5).

Discussion

In this meta-analysis, FDG-PET was considered positive in 142 of 225 patients (63%) with a thyroid nodule with either indeterminate or repeatedly inconclusive cytology of FNAB. When surgery was performed consequently, almost 40% of these were confirmed at final histopathology as carcinoma. A negative FDG-PET demonstrated malignancy in only 3 of 83 (3.6%) of these individual patients without focal FDG-uptake, all having a histological diameter below 1.5 cm. In the clinical work-up for patients with larger nodules FDG-PET is considered a useful tool and can reliably exclude T2 malignancies by a negative FDG-PET.

We computed that it would require 72 patients with thyroid malignancy all having a positive FDG-PET scan to be reasonably confident the true sensitivity is higher than 95% (i.e. the lower level of the 2-tailed Clopper-Pearson 95-CI or the solution to $1 - \beta_{1-\frac{\alpha}{2}}(n - x + 1, x) \geq 0.95$ with n=x=1 and α=0.05). Since the prevalence of malignancy in TN in the presented data was 25.8%, at least 276 patients with TN with indeterminate FNAB should have been included. Up to date, published studies have sample sizes much smaller than this number, therefore this comprehensive meta-analysis, was undertaken.

As meta-analyses are prone to error due to factors such as low study quality, study inhomogeneity and publication bias these factors were minimized by careful selection and quality appraisal of references and description of potential causes of heterogeneity and publication bias. This resulted in a pooled LR- of 0.19 (fixed effects) - 0.24 (random effects), indicating that the pre-FDG-PET probability of malignancy in these patients decreases from 26% (the prevalence) to 6.2% (fixed effects) or 7.7% (random effects) after a negative FDG-PET scan. In other words, incorporation of FDG-PET in the workup of a TN with indeterminate FNAB rightly saves 27 patients from diagnostic surgery at the cost of one patient that is unjustifiably delayed surgical treatment (80/225 patients would rightly not be operated upon when incorporating FDG-PET in the work-up of a TN (true negatives), however 3/225 patients would faultily not be operated upon due to a false negative result: 80/3≈27/1). The number of false negatives can be decreased by only taking lesions larger than 15mm

by pathology in diameter into consideration. This might be explained as a consequence of the partial-volume effect due to the limited spatial resolution (>~6mm) of used PET scanners.

The total costs for FDG-PET in The Netherlands currently is approximately €1,400 (in the US, for comparison the reimbursement rate currently is approximately $1050) and the treatment costs are mainly driven by the costs of surgery and hospitalization, with mean costs per patient amounting to €3,311 in benign and €5,228 in malignant cases, without considering additional-treatment costs, economic costs or indirect costs.[42, 43] In the pooled population in this meta-analysis, surgery in all patients would result in an average cost per patients of €3,805 without the use of FDG-PET and € 3958 with the use of FDG-PET in the Netherlands (approximately €3,608 in based on the reimbursement rate in the US). Thus, both scenarios generate similar direct costs, while additional-treatment costs, economic costs or indirect costs of futile surgery are not even considered. For this computation, we did not consider extra costs for the 1.3% false-negative FDG-PET scans.

However, there remain other reasons to consider surgery despite this cost-effectiveness: mechanical or cosmetic concerns or mere reassurance. Nonetheless, as in larger lesions the complication rates can be higher (due to extension towards the large vessels, trachea or recurrent laryngeal nerve) these patients might benefit most of reassurance by a true-negative PET. However in patients with small lesions (<15mm) the value of FDG-PET should be weighed against the disadvantage of the risk of a false-negative result. When renouncing surgery in these patients, follow-up remains warranted.

A limitation of this meta-analysis is that different studies used different definitions of FDG-PET-positivity. For example, Hales *et al.*[38] required a threshold SUV_{max} of 2.0 for a positive test. It was tried to use exactly the same definition for FDG-PET-positivity for each reference. However, since the FDG-PET scans were not examined centrally, still some inhomogeneity due to inter-observer variability in interpretation of these images remain. Another limitation is the definition of malignancy: some excluded incidentally found papillary microcarcinomas from participation in the study[36, 37], others even considered malignant lesions found distant from the (benign) index nodule.[41] Of the four lesions of latter category (0.3, 4, 8 and 17mm PTC), two were FDG-PET negative and could therefore also be considered false-negative rather than true-negative scans. This would have increased the total

number of false negative scans to 5 and therefore decreased the pooled sensitivity from 95% (95-CI: 86-99%) to 92% (95-CI: 82-97%). A final limitation is concerning the population heterogeneity, particularly the vast variation in prevalence of malignancy (14-42%) in different parts of the world with both endemic goiter and iodine-sufficient areas.

Conclusions

This comprehensive systematic review and meta-analysis of the literature showed that in patients with TN with indeterminate FNAB, a negative FDG-PET scan improves diagnostic accuracy, particularly in patients with lesions larger than 15 mm. All false-negatives FDG-PET cases were lesions smaller than 15 mm (i.e. T1 tumors). A positive FDG-PET increases the chance of malignancy from 25.8% to 38.7% in these patients. Further prospective series are ongoing and will ultimately reveal the value of FDG-PET in the diagnostic evaluation of thyroid nodules to confirm these findings.

References

1. Altekruse SF, Kosary CL, Krapcho M, Neyman N, Aminou R, Waldron W, et al. *SEER Cancer Statistics Review, 1975-2007; http://seer.cancer.gov/csr/1975_2007/.* National Cancer Institute. Bethesda, MD, based on November 2009 SEER data submission, posted to the SEER web site, 2010.

2. Netea-Maier RT, Aben KK, Casparie MK, den Heijer M, Grefte JM, Slootweg P, et al. Trends in incidence and mortality of thyroid carcinoma in The Netherlands between 1989 and 2003: correlation with thyroid fine-needle aspiration cytology and thyroid surgery. *Int J Cancer* 2008;123(7):1681-4.

3. Sipos JA, Mazzaferri EL. Thyroid cancer epidemiology and prognostic variables. *Clin Oncol (R Coll Radiol)* 2010;22(6):395-404.

4. Ezzat S, Sarti DA, Cain DR, Braunstein GD. Thyroid incidentalomas. Prevalence by palpation and ultrasonography. *Arch Intern Med* 1994;154(16):1838-40.

5. Cooper DS, Doherty GM, Haugen BR, Kloos RT, Lee SL, Mandel SJ, et al. Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. *Thyroid* 2009;19(11):1167-214.

6. Tan GH, Gharib H. Thyroid incidentalomas: management approaches to nonpalpable nodules discovered incidentally on thyroid imaging. *Ann Intern Med* 1997;126(3):226-31.

7. Hegedus L. Clinical practice. The thyroid nodule. *N Engl J Med* 2004;351(17):1764-71.

8. Baloch ZW, LiVolsi VA, Asa SL, Rosai J, Merino MJ, Randolph G, et al. Diagnostic terminology and morphologic criteria for cytologic diagnosis of thyroid lesions: a synopsis of the National Cancer Institute Thyroid Fine-Needle Aspiration State of the Science Conference. *Diagn Cytopathol* 2008;36(6):425-37.

9. Carpi A, Mechanick JI, Saussez S, Nicolini A. Thyroid tumor marker genomics and proteomics: diagnostic and clinical implications. *J Cell Physiol* 2010;224(3):612-9.

10. Buerkle A, Weber WA. Imaging of tumor glucose utilization with positron emission tomography. *Cancer Metastasis Rev* 2008;27(4):545-54.

11. Cohen MS, Arslan N, Dehdashti F, Doherty GM, Lairmore TC, Brunt LM, et al. Risk of malignancy in thyroid incidentalomas identified by fluorodeoxyglucose-positron emission tomography. *Surgery* 2001;130(6):941-6.

12. Chen YK, Ding HJ, Chen KT, Chen YL, Liao AC, Shen YY, et al. Prevalence and risk of cancer of focal thyroid incidentaloma identified by 18F-fluorodeoxyglucose positron emission tomography for cancer screening in healthy subjects. *Anticancer Research* 2005;25(2B):1421-6.

13. Kang KW, Kim SK, Kang HS, Lee ES, Sim JS, Lee IG, et al. Prevalence and risk of cancer of focal thyroid incidentaloma identified by 18F-fluorodeoxyglucose positron emission tomography for metastasis evaluation and cancer screening in healthy subjects. *J Clin Endocrinol Metab* 2003;88(9):4100-4.

14. Shie P, Cardarelli R, Sprawls K, Fulda KG, Taur A. Systematic review: prevalence of malignant incidental thyroid nodules identified on fluorine-18 fluorodeoxyglucose positron emission tomography. *Nucl Med Commun* 2009;30(9):742-8.

15. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6(7):e1000097.

16. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med* 2009;6(7):e1000100.

17. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. *ACP J Club* 1995;123(3):A12-3.

18. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;326(7379):41-4.

19. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem* 2003;49(1):7-18.

20. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;3:25.

21. Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol* 2006;6:9.

22. Vriens D, Visser EP, de Geus-Oei LF, Oyen WJ. Methodological considerations in quantification of oncological FDG PET studies. *Eur J Nucl Med Mol Imaging* 2010;37(7):1408-25.

23. de Geus-Oei LF, Pieters GF, Bonenkamp JJ, Mudde AH, Bleeker-Rovers CP, Corstens FH, et al. 18F-FDG PET reduces unnecessary hemithyroidectomies for thyroid nodules with inconclusive cytologic results. *J Nucl Med* 2006;47(5):770-5.

24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33(1):159-74.

25. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;56(11):1129-35.

26. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med* 1998;17(8):857-72.

27. Jones CM, Athanasiou T. Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests. *Ann Thorac Surg* 2005;79(1):16-20.

28. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 1993;12(14):1293-316.

29. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21(11):1539-58.

30. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol* 2005;58(9):882-93.

31. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol* 2000;53(11):1119-29.

32. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315(7109):629-34.

33. Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. *BMC Med Res Methodol* 2006;6:31.

34. Kresnik E, Gallowitsch HJ, Mikosch P, Stettner H, Igerc I, Gomez I, et al. Fluorine-18-fluorodeoxyglucose positron emission tomography in the preoperative assessment of thyroid nodules in an endemic goiter area. *Surgery* 2003;133(3):294-9.

35. Mitchell JC, Grant F, Evenson AR, Parker JA, Hasselgren PO, Parangi S. Preoperative evaluation of thyroid nodules with 18FDG-PET/CT. *Surgery* 2005;138(6):1166-74; discussion 74-5.

36. Kim JM, Ryu J-S, Kim TY, Kim WB, Kwon GY, Gong G, et al. 18F-fluorodeoxyglucose positron emission tomography does not predict malignancy in thyroid nodules cytologically diagnosed as follicular neoplasm. *Journal of Clinical Endocrinology & Metabolism* 2007;92(5):1630-4.

37. Sebastianes FM, Cerci JJ, Zanoni PH, Soares J, Jr., Chibana LK, Tomimori EK, et al. Role of 18F-fluorodeoxyglucose positron emission tomography in preoperative assessment of cytologically indeterminate thyroid nodules.[Erratum appears in J Clin Endocrinol Metab. 2008 Jan;93(1):81]. *Journal of Clinical Endocrinology & Metabolism* 2007;92(11):4485-8.

38. Hales NW, Krempl GA, Medina JE. Is there a role for fluorodeoxyglucose positron emission tomography/computed tomography in cytologically indeterminate thyroid nodules? *Am J Otolaryngol* 2008;29(2):113-8.

39. Smith RB, Robinson RA, Hoffman HT, Graham MM. Preoperative FDG-PET imaging to assess the malignant potential of follicular neoplasms of the thyroid. *Otolaryngol Head Neck Surg* 2008;138(1):101-6.

40. D'Souza MM, Marwaha RK, Sharma R, Jaimini A, Thomas S, Singh D, et al. Prospective evaluation of solitary thyroid nodule on 18F-FDG PET/CT and high-resolution ultrasonography. *Ann Nucl Med* 2010.

41. Traugott AL, Dehdashti F, Trinkaus K, Cohen M, Fialkowski E, Quayle F, et al. Exclusion of malignancy in thyroid nodules with indeterminate fine-needle aspiration cytology after negative 18F-fluorodeoxyglucose positron emission tomography: interim analysis. *World J Surg* 2010;34(6):1247-53.

42. Krug B, Van Zanten A, Pirson AS, Crott R, Borght TV. Activity-based costing evaluation of a [(18)F]-fludeoxyglucose positron emission tomography study. *Health Policy* 2009;92(2-3):234-43.

43. Hooft L, Hoekstra OS, Boers M, Van Tulder MW, Van Diest P, Lips P. Practice, efficacy, and costs of thyroid nodule evaluation: a retrospective study in a Dutch university hospital. *Thyroid* 2004;14(4):287-93.

Tables

Ref.:	First author (year):	City / country:	Hospital type:	Inclusion criteria:	Exclusion criteria:	TSH:	STN:	Remarks:
[34]	Kresnik (2003)	Klagenfurt, Austria	G	- All patients with TN - hypoechogenic or no uptake on scintigraphy - FNAB: follicular or Hürthle cell proliferation - scheduled for surgery	- Autonomous goiter	Normal in all	24/43	Endemic goiter area
[23]	De Geus-Oei (2006)	Nijmegen, the Netherlands	U&G	- palpable TN - inconclusive FNAB - scheduled for hemithyroidectomy	- DM - pregnancy	Normal in all	44/44	Multicentre trial
[36]	Kim (2007)	Seoul, South-Korea	U	- TN larger than 1 cm - FNAB: follicular proliferation		Normal in all	32/46	
[37]	Sebastianes (2007)	São Paulo, Brazil	U	- FNAB: indeterminate - scheduled for (hemi)thyroidectomy	- uncontrollable DM - other malignancies - pregnancy - abnormal TSH	Normal in all	21/42	
[38]	Hales (2008)	Oklahoma City (OK), USA	U	- All patients with TN - FNAB: indeterminate	- pregnancy - breastfeeding - prior H&N surgery - >181 kg bodyweight	?	8/15	
[41]	Traugott (2010)	St Louis (MO), USA	U	- Adults with TN or dominant TN - palpable or >1cm on US - scheduled for surgery	- prior neck surgery - prior radiotherapy	Normal in all	51/51	Interim Analysis

Table 1: Characteristics of the patient population included in the studies selected for this meta-analysis. DM: Diabetes Mellitus, FNAB: Fine Needle Aspiration Biopsy; G: General hospital; H&N: Head and Neck; STN: Single Thyroid Nodule; U: University Hospital; US: Ultrasonography.

Ref.:	First author (year):	N	Female [%]:	Mean age [yr]: (±SD)	Median hist. lesion diameter [mm]: (range)	FNAB:				
						Follicular proliferation:	Hürthle cell proliferation:	Combination:	Repetitively inconclusive:	Indeterminate NOS:
[34]	Kresnik (2003)*	37	78%	55.1 (±13.8)	16 (6-80)	65%	30%	-	5%	-
[23]	De Geus-Oei (2006)	44	93%	48.5 (±13.8)	30# (3-55)	75%	9%	11%	5%	-
[36]	Kim (2007)*	36	86%	44.2 (±13.1)	25 (10-90)	100%	-	-	-	-
[37]	Sebastianes (2007)	42	90%	45.3 (±16.3)	28 (4-85)	-	-	-	-	100%
[38]	Hales (2008)	15	93%	47.5 (±14.9)	25 (1-60)	80%	13%	7%	-	-
[41]	Traugott (2010)	51	80%	49.6 (±10.6)	15 (5-50)¶	69%	10%	-	-	21%
Test for between studies differences			**p=0.292†**	**p=0.012‡**	**p<0.001****	**p<0.001†**				
Total		**225**	**86%**	**48.5 (±13.9)**	**20 (15-35)**	**62%**	**10%**	**3%**	**2%**	**23%**

Table 2: Clinical characteristics of the patients included in the meta-analysis (extracted data). FNAB: Fine Needle Aspiration Biopsy; Hist.: histology; SD: standard deviation; NOS: Not Otherwise Specified; *Presented data varies from published data: not all patients used for publications could be used for meta-analysis; #Lesion size was unavailable in 9 lesions; ¶ Lesion size was unavailable in 1 lesion; †Computed using Fisher's Exact (FE) test; ‡Computed using one-way ANOVA, difference caused by older patients in study of Kresnik (2003)[34] compared to Kim (2007)[36] and Sebastianes (2007)[37] (Tukey HSD, p=0.010 and p=0.020, respectively); **Computed using Kruskal-Wallis H.

Ref.:	First Author (year):	TP	TN	FP	FN	Prev	Se	Sp	NPV	PPV	Acc
[34]	Kresnik (2003)*	10	15	12	0	27.0% [13.8-44.1%]	100% [69.2-100%]	55.6% [35.3-74.5%]	100% [78.2-100%]	45.5% [24.4-67.8%]	67.6% [50.2-82.0%]
[23]	De Geus-Oei (2006)	6	25	13	0	13.6% [5.2-27.4%]	100% [54.1-100%]	65.8% [48.6-80.4%]	100% [86.3-100%]	31.6% [12.6-56.6%]	70.5% [54.8-83.2%]
[36]	Kim (2007)*	15	0	21	0	41.7% [25.5-59.2%]	100% [78.2-100%]	0% [0-16.1%]	NaN	41.7% [25.5-59.2%]	41.7% [25.5-59.2%]
[37]	Sebastianes (2007)	11	12	19	0	26.2% [13.9-42.0%]	100% [71.5-100%]	38.7% [21.8-57.8%]	100% [73.5-100%]	36.7% [19.9-56.1%]	54.8% [38.7-70.2%]
[38]	Hales (2008)	5	3	6	1	40.0% [16.3-67.7%]	83.3% [35.9-99.6%]	33.3% [7.5-70.1%]	75.0% [19.4-99.4%]	45.5% [16.7-76.6%]	53.3% [26.6-78.7%]
[41]	Traugott (2010)	8	25	16	2	19.6% [9.8-33.1%]	80.0% [44.4-97.5%]	61.0% [44.5-75.8%]	92.6% [75.7-99.1%]	33.3% [15.6-55.3%]	64.7% [50.1-77.6%]
Pooled		**55**	**80**	**87**	**3**	**25.8% [20.2-32.0%]**	**94.8% [85.6-98.9%]**	**47.9% [40.1-55.8%]**	**96.4% [89.9-99.2%]**	**38.7% [30.7-47.3%]**	**60.0% [53.3-66.5%]**

Table 3: Test characteristics of FDG-PET [with 95-confidence interval] for detection of malignancy in TN with indeterminate FNAB according to references. Acc: accuracy; FN: false negatives; FP: false positives; NaN: not a number; NPV: negative predictive value; TN: true negatives; TP: true positives; PPV: positive predictive value; Prev: cancer prevalence; Se: sensitivity; Sp: specificity

Figure Captions

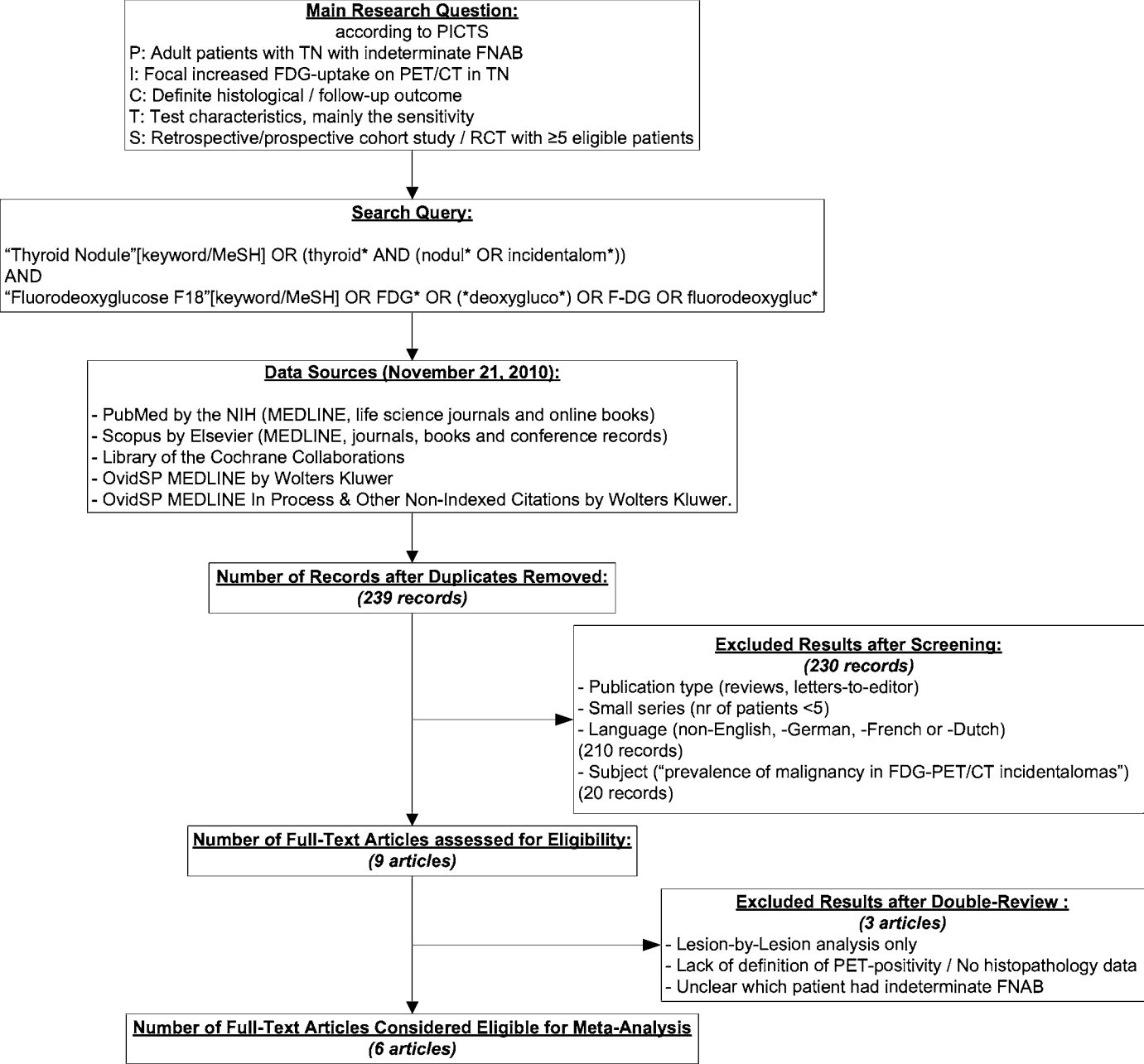
Figure 1: Flowchart of selection of articles. FDG: [^{18}F]-2-fluoro-2-deoxy-D-glucose; FNAB: Fine-Needle Aspiration Biopsy; MeSH: Medical Subject Heading; PET/CT: Positron Emission Tomography / Computed Tomography; PICTS: target Population, Index test, Comparator test, Target condition and Study design; RCT: Randomized Controlled Trial; TN: Thyroid Nodule; * marks truncation.

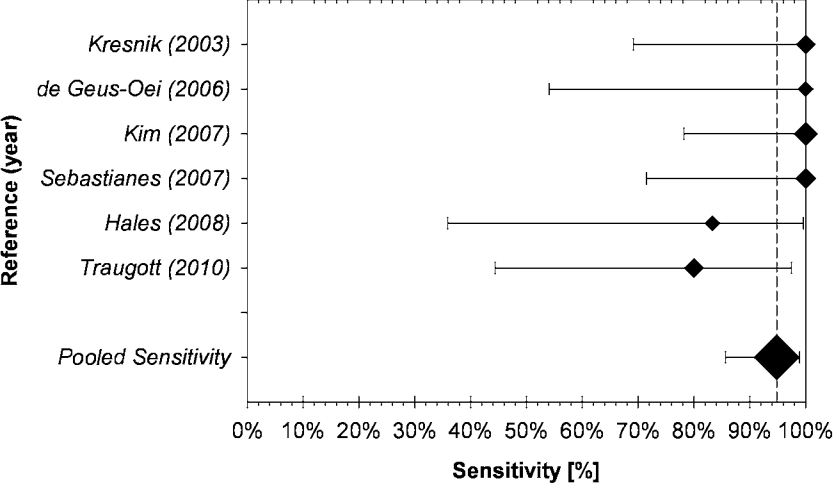
Figure 2: Forest plots of sensitivity, specificity, positive and negative likelihood ratio with corresponding 95-confidence intervals of FDG-PET in detection of malignancy in TN with indeterminate FNAB. Sensitivity and specificity are pooled using individual patient data. The likelihood ratios are pooled using either fixed effects (FE) or random effects (RE) modeling. The dashed lines represent the pooled effects. The dotted lines represent a likelihood ratio of 1 (i.e. no change in likelihood).

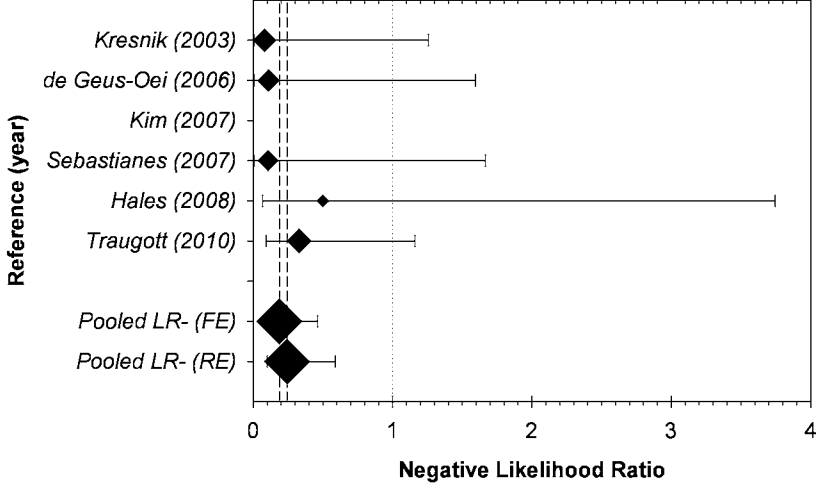
Figure 3: Summary Receiver Operating Characteristic (sROC) curve of sensitivity versus 1-specificity of FDG-PET in detection of malignancy in TN with indeterminate FNAB with corresponding boundaries of the 95-confidence interval. There are no signs of threshold effects. The area under the curve is 0.84 (±0.079).

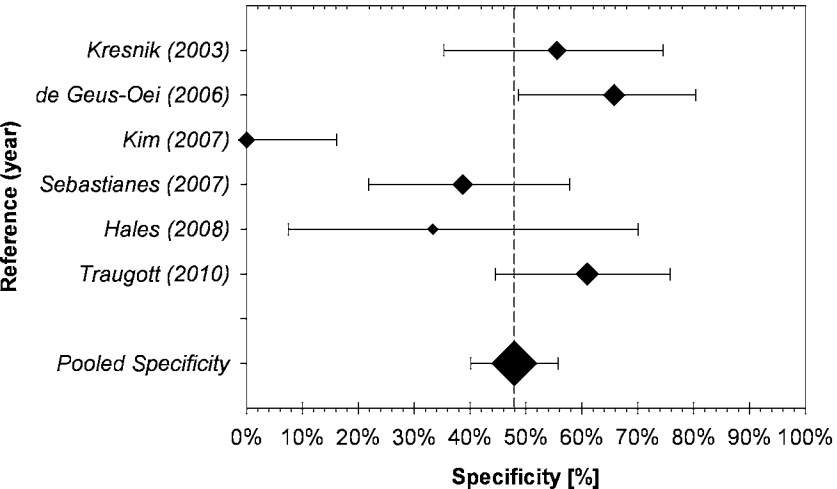
Figure 4: Funnel plot of the diagnostic Odds Ratios (dOR) of included references. The solid vertical line denotes the pooled dOR and the dashed lines the 95% confidence intervals (95-CI). The standard error (SE) of de dOR as well as the pooled dOR was calculated using Mantel-Haenzels weighing (fixed effects, FE). Symmetric distribution of references can be seen. No correlation could be detected between $\log_e$(dOR) and SE($\log_e$(dOR)) as reflected by Kendall τ_b=0.067 (p=0.851), therefore there is no evidence for publication bias.

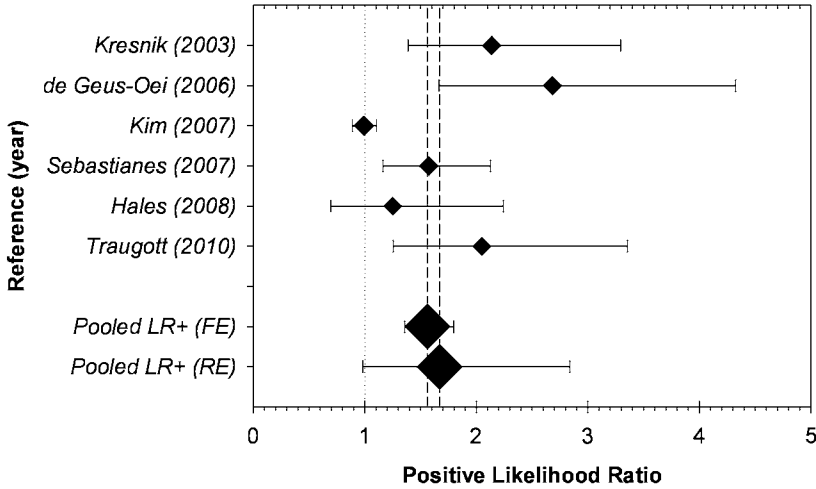
Figure 5: Dependency of sensitivity (Se) and specificity (Sp) with 95-confidence intervals of FDG-PET in detection of malignancy in TN with indeterminate FNAB. *n*: the number of patients for which the diagnostic test characteristics are computed.

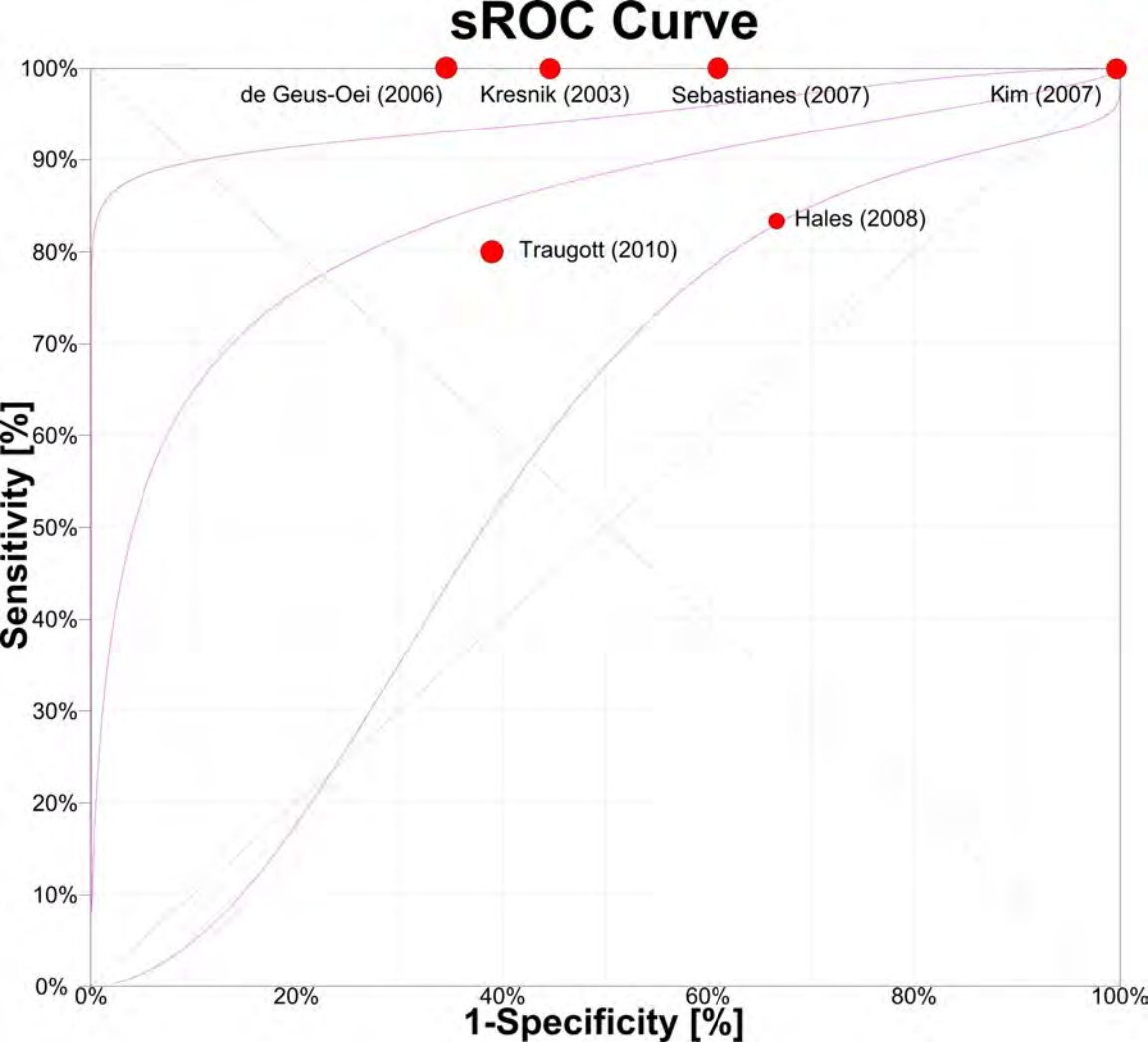

Main Research Question:
according to PICTS
P: Adult patients with TN with indeterminate FNAB
I: Focal increased FDG-uptake on PET/CT in TN
C: Definite histological / follow-up outcome
T: Test characteristics, mainly the sensitivity
S: Retrospective/prospective cohort study / RCT with ≥5 eligible patients
Search Query:
"Thyroid Nodule"[keyword/MeSH] OR (thyroid* AND (nodul* OR incidentalom*))
AND
"Fluorodeoxyglucose F18"[keyword/MeSH] OR FDG* OR (*deoxygluco*) OR F-DG OR fluorodeoxygluc*
Data Sources (November 21, 2010):
- PubMed by the NIH (MEDLINE, life science journals and online books)
- Scopus by Elsevier (MEDLINE, journals, books and conference records)
- Library of the Cochrane Collaborations
- OvidSP MEDLINE by Wolters Kluwer
- OvidSP MEDLINE In Process & Other Non-Indexed Citations by Wolters Kluwer.
Number of Records after Duplicates Removed:
(239 records)
Excluded Results after Screening:
(230 records)
- Publication type (reviews, letters-to-editor)
- Small series (nr of patients <5)
- Language (non-English, -German, -French or -Dutch)
(210 records)
- Subject ("prevalence of malignancy in FDG-PET/CT incidentalomas")
(20 records)
Number of Full-Text Articles assessed for Eligibility:
(9 articles)
Excluded Results after Double-Review :
(3 articles)
- Lesion-by-Lesion analysis only
- Lack of definition of PET-positivity / No histopathology data
- Unclear which patient had indeterminate FNAB
Number of Full-Text Articles Considered Eligible for Meta-Analysis
(6 articles)

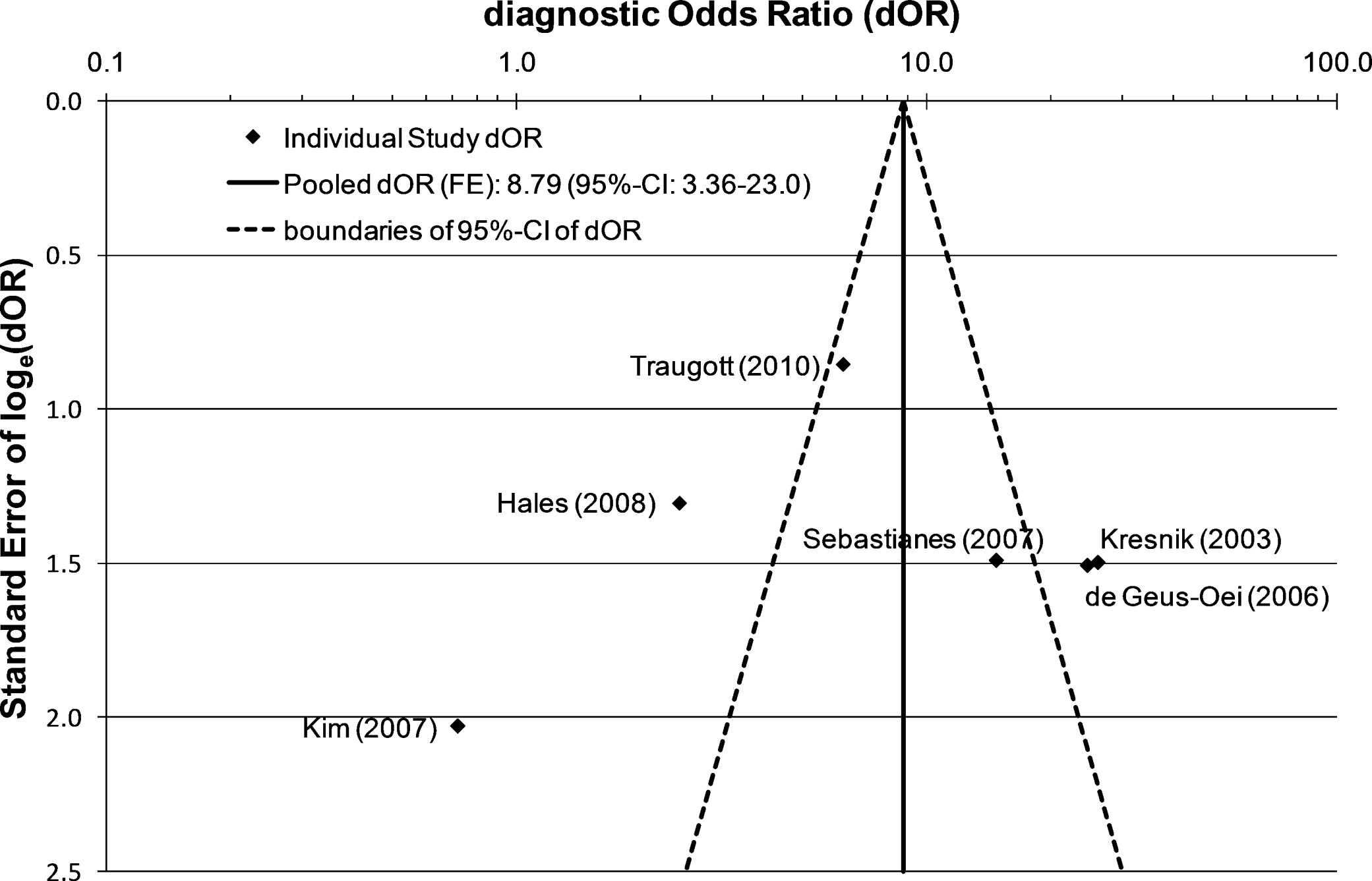

Kresnik (2003)
de Geus-Oei (2006)
Kim (2007)
Sebastianes (2007)
Hales (2008)
Traugott (2010)
Pooled Sensitivity
Reference (year)
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Sensitivity [%]

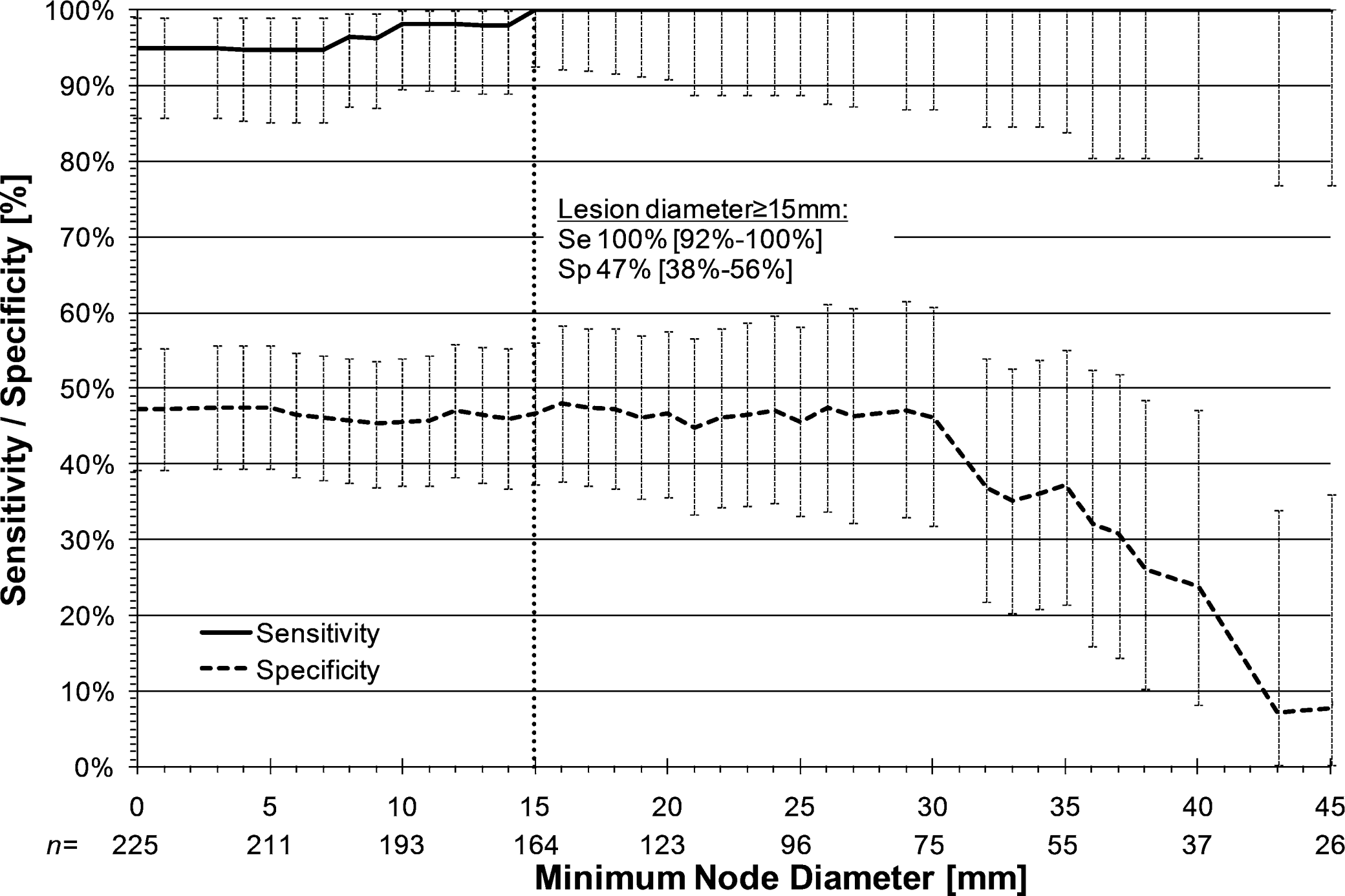

Kresnik (2003)
de Geus-Oei (2006)
Kim (2007)
Sebastianes (2007)
Hales (2008)
Traugott (2010)
Pooled LR- (FE)
Pooled LR- (RE)
Reference (year)
0 1 2 3 4
Negative Likelihood Ratio


Kresnik (2003)
de Geus-Oei (2006)
Kim (2007)
Sebastianes (2007)
Hales (2008)
Traugott (2010)
Pooled Specificity
Reference (year)
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Specificity [%]


Kresnik (2003)
de Geus-Oei (2006)
Kim (2007)
Sebastianes (2007)
Hales (2008)
Traugott (2010)
Pooled LR+ (FE)
Pooled LR+ (RE)
Reference (year)
0 1 2 3 4 5
Positive Likelihood Ratio


sROC Curve
de Geus-Oei (2006)
Kresnik (2003)
Sebastianes (2007)
Kim (2007)
Hales (2008)
Traugott (2010)
Sensitivity [%]
1-Specificity [%]
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
20%
40%
60%
80%
100%


diagnostic Odds Ratio (dOR)
0.1
1.0
10.0
100.0
0.0
0.5
1.0
1.5
2.0
2.5
Standard Error of log$_e$(dOR)
Individual Study dOR
Pooled dOR (FE): 8.79 (95%-CI: 3.36-23.0)
boundaries of 95%-CI of dOR
Traugott (2010)
Hales (2008)
Sebastianes (2007)
Kresnik (2003)
de Geus-Oei (2006)
Kim (2007)


Sensitivity / Specificity [%]
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Lesion diameter≥15mm:
Se 100% [92%-100%]
Sp 47% [38%-56%]
Sensitivity
Specificity
0
5
10
15
20
25
30
35
40
45
n=
225
211
193
164
123
96
75
55
37
26
Minimum Node Diameter [mm]