HHS Public Access

Author manuscript

Ann Surg Oncol. Author manuscript; available in PMC 2016 June 28.

Published in final edited form as:

Ann Surg Oncol. 2015 November ; 22(12): 3996–4001. doi:10.1245/s10434-015-4486-3.

Wide Inter-institutional Variation in Performance of a Molecular Classifier for Indeterminate Thyroid Nodules

Jennifer L. Marti, MD, FACS[1], **Vaidehi Avadhani, MD**[2], **Luke A. Donatelli, MD**[3], **Sayani Niyogi, DO**[3], **Beverly Wang, MD**[2], **Richard J. Wong, MD, FACS**[3], **Ashok R. Shaha, MD, FACS**[3], **Ronald A. Ghossein, MD**[4], **Oscar Lin, MD**[4], **Luc G. T. Morris, MD, MSc, FACS**[3], and **Allen S. Ho, MD**[3]

Luc G. T. Morris: morrisl@mskcc.org; Allen S. Ho: allen.ho@cshs.org

[1]Division of Endocrine Surgery, Department of Surgery, Mount Sinai Beth Israel, Icahn School of Medicine at Mount Sinai, New York, NY

[2]Department of Pathology, Mount Sinai Beth Israel, Icahn School of Medicine at Mount Sinai, New York, NY

[3]Head and Neck Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY

[4]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY

Abstract

Background—The Afirma gene expression classifier (GEC) is used to assess malignancy risk in indeterminate thyroid nodules (ITNs) classified as Bethesda category III/IV. Our objective was to analyze GEC performance at two institutions with high thyroid cytopathology volumes but differing prevalence of malignancy.

Methods—Retrospective analysis of all ITNs evaluated with the GEC at Memorial Sloan Kettering Cancer Center (MSK; $n = 94$) and Mount Sinai Beth Israel (MSBI; $n = 71$). These institutions have differing prevalences of malignancy in ITNs: 30–38 % (MSK) and 10–19 % (MSBI). Surgical pathology was correlated with GEC findings for each matched nodule. Performance characteristics were estimated using Bayes Theorem.

Results—Patient and nodule characteristics were similar at MSK and MSBI. The GEC-benign call rates were 38.3 % (MSK) and 52.1 % (MSBI). Of the GEC-benign nodules, 8.3 % (MSK) and 13.5 % (MSBI) were treated surgically. Surgical pathology indicated that all of GEC-benign nodules were benign. Of the GEC-suspicious nodules, 60.0 % (MSK) and 61.7 % (MSBI) underwent surgery. Positive predictive values (PPVs) for GEC-suspicious results were 57.1 % (95 % CI 41.0–72.3) at MSK and 14.3 % (95 % CI 0.2–30.2) at MSBI. The estimated negative predictive values (NPVs) were 86–92 % at MSK and 95–98 % at MSBI.

Conclusions—There were wide variations in the Afirma GEC-benign call rate, PPV, and NPV between MSBI (a comprehensive health system) and MSK (a tertiary referral cancer center),

---

This research was delivered as a podium presentation at the 2015 Society of Surgical Oncology meeting.

DISCLOSURE The authors have no potential conflict of interest

which had differing rates of malignancy in ITNs. The GEC could not routinely alter management in either institution. We believe that this assay would be expected to be most informative in practice settings where the prevalence of malignancy is 15–21 %, such that NPV >95 % and PPV >25 % would be anticipated. Knowing the prevalence of malignancy in ITNs at a particular institution is critical for reliable interpretation of GEC results.

Despite advances in ultrasonography and fine-needle aspiration (FNA), 10–30 % of biopsied thyroid nodules have cytopathology that is categorized as “indeterminate.” Indeterminate thyroid nodules (ITNs) are defined, using the Bethesda System for Reporting Thyroid Cytopathology, as Category III (atypia of undetermined significance or follicular lesion of undetermined significance) or Category IV (suspicious for follicular neoplasm).[1] The most frequently recommended approach, diagnostic thyroid lobectomy, reflects the practical limitations of diagnosing thyroid nodules. If the nodule is malignant, the patient often requires another operation for completion thyroidectomy. If it is benign, the surgery in retrospect may have been unnecessary.

The Afirma gene expression classifier (GEC) is a molecular assay that was developed for the purpose of improving surgical decision-making for ITNs classified as Bethesda III or IV.[2] The GEC categorizes ITNs as either “benign” or “suspicious.” Results of a prospective, multicenter, validation study demonstrated the assay’s ability to correctly identify the majority of benign thyroid nodules, even when cytologically indeterminate, with sensitivity of 90 % and specificity of 49–53 %. In the multicenter trial, this test was able to achieve a negative predictive value (NPV) of 94–95 % and a positive predictive value (PPV) of 37–38 %.

However, PPV and NPV values are not absolute. For any diagnostic test with a given sensitivity and specificity, the NPV and PPV depend heavily on the pre-test prevalence of malignancy in the population being studied. As the prevalence of malignancy rises, the PPV increases and the NPV decreases. Thus, GEC performance could range widely depending on an institution’s practice and the patients’ characteristics. Consequently, GEC performance in a prospective trial may not be generalizable to all practice settings. In the present study, the performance of GEC at two institutions—each with a high volume of thyroid cytopathology but with differing prevalences of malignancy—was evaluated.

MATERIALS AND METHODS

We performed a retrospective analysis of all ITNs evaluated with the Afirma GEC assay at Memorial Sloan Kettering Cancer Center (MSK) (n = 94 nodules) or Mount Sinai Beth Israel (MSBI) (n = 71 nodules) in New York City between February 2013 and December 2014. Both centers have a high volume of patients with thyroid nodules and thyroid cancer, but have differing scopes of practice and thyroid patient referral patterns. MSK is a tertiary referral cancer center, whereas MSBI operates as a comprehensive health care system. Of the 94 MSK patients (94 nodules), 47 were internal patients with Afirma testing performed at MSK, and 47 were external patients with Afirma testing already performed elsewhere. All 62 MSBI patients (with 71 nodules) were internal cases. We previously analyzed ITNs at

each institution and reported the prevalence of malignancy in ITNs as 30–38 % at MSK and 10–19 % at MSBI.

All reports followed the diagnostic scheme proposed by the Bethesda System for Reporting Thyroid Cytopathology, and were generated by dedicated, fellowship-trained thyroid cytopathologists at each institution. Surgical pathology results for each nodule were correlated with FNA/GEC findings by matching the biopsied nodule to the resected nodule. All incidental carcinomas observed elsewhere in the thyroid gland were excluded when calculating the malignancy rates. All patients with additional nodules harboring nodules that were categorized as Bethesda V (suspicious for malignancy) or Bethesda VI (malignant) were excluded, as were those without documented follow-up.

These data were used to calculate the GEC-benign call rate (the percentage of GEC tests that resulted in a “benign” diagnosis) and the PPV (true positives divided by the number of GEC-suspicious results). The NPV could not be calculated directly because the GEC-benign nodules generally were not treated surgically. The NPV, PPV, and benign call rate were all modeled via Bayes Theorem, permitting comparison of the anticipated and the observed benign call rate and the PPV as well as estimation of the expected NPV. Bayes Theorem is a technique that allows one to calculate the probability of disease given the results of an imperfect test. This analysis incorporates the pre-test probability of disease (e.g., the prevalence of cancer in ITNs) and the characteristics of the test (sensitivity and specificity). If these values are known, Bayes Theorem permits calculation of the probability of disease given a positive or a negative test result. These numbers are the PPV and NPV of a diagnostic test.

Sensitivity and specificity data were based on figures reported by Alexander et al. To avoid bias from patients specifically referred to an institution for surgery for a GEC-suspicious result, only nodules worked up internally were used to calculate the benign call rate. Confidence intervals were calculated based on a Poisson distribution. This study was reviewed and deemed exempt by both the MSK and MSBI institutional review boards.

RESULTS

Trends in the estimated performance characteristics of the GEC (benign call rate, PPV, NPV) based on pre-test probability (the prevalence of malignancy in ITNs) and test characteristics (sensitivity and specificity) are shown in Fig. 1. Other patient and nodule characteristics were similar at the two institutions (Table 1).

In all, 94 patients, each with one ITN, were identified at MSK. Among the internally worked up (non referral) MSK patients, 18 had a GEC-benign nodule and 29 had a GEC-suspicious nodule. At MSBI, 62 patients (71 ITNs) were identified during the study period. Among these 62 patients, 29 had one GEC-benign nodule, 24 had one GEC-suspicious nodule, and 9 had two ITNs that were sent for GEC evaluation: two patients with two GEC-suspicious nodules, one patient with two GEC-benign nodules, and six patients each with one GEC-benign and one GEC-suspicious nodule.

Cytological evaluation showed that 67.0 % (MSK) and 56.3 % (MSBI) were Bethesda III nodules. Among the surgical cases that were found to be malignant, 62.5 % (15/24) at MSK were a follicular variant of papillary thyroid carcinoma, whereas 33.3 % (1/3) were this type of carcinoma at MSBI (Fig. 2).

The GEC-benign call rate was 38.3 % at MSK compared with 52.1 % at MSBI. Among the GEC-benign nodules, 8.3 % (2/24) at MSK and 13.5 % (5/37) at MSBI underwent surgery despite the benign GEC result. Indications for surgery were enlarging nodule size, compressive symptoms, or an additional nodule that was GEC-suspicious. In both institutions, all resected GEC-benign nodules were declared benign based on their surgical pathology. Among the GEC-suspicious nodules, 60.0 % (42/70) at MSK and 61.7 % (21/34) at MSBI had undergone surgery at the time of analysis (Fig. 2).

After matching GEC data of the biopsied nodule to the final surgical pathology of the same nodule, the PPV of a GEC-suspicious result was 57.1 % [95 % confidence interval (CI) 41.0–72.3] at MSK and 14.3 % (95 % CI 4.1–35.5) at MSBI (Fig. 2). These numbers do not include incidental carcinomas separate from the biopsied nodule (among the resected glands with benign biopsied nodules, there were eight incidental carcinomas at MSK and two at MSBI). The estimated NPVs were 86–92 % (MSK) and 95–98 % (MSBI). The observed benign call rates and PPVs were similar to those predicted by Bayes Theorem (Fig. 1). Based on the known prevalence of malignancy at MSK, the expected benign call rate was 35–38 % (actual: 38.3 %), and the estimated PPV was 44–55 % (actual 57.1 %). At MSBI, the estimated benign call rate was 43–47 % (actual 52.1 %), and the estimated PPV was 15–30 % (actual 14.3 %). The wide variance in GEC performance depended on the pre-test probability of malignancy, as shown in Fig. 3.

DISCUSSION

This study characterized the variable clinical performance of the Afirma gene expression-based molecular assay for determining the risk of cancer in ITNs. We compared the performance of this test at two institutions in New York City that have high-volume thyroid cytopathology programs but differing patient referral patterns—and therefore differing prevalences of malignancy in ITNs. At MSK, a cancer referral center, the prevalence of malignancy was approximately double that at MSBI, which is a comprehensive health care system. We showed that this difference leads to wide variation in the performance of the GEC test for ITNs.

Comparing data from the two institutions, we found wide variation in the benign call rate, PPV, and NPV of the GEC assay. We found that the observed benign call rate and PPV figures varied widely between the two institutions but were in line with predicted values (Fig. 1). At both institutions, the GEC results did not routinely alter management of ITNs. At MSBI, where the pre-test probability of an ITN being cancer was low, the GEC results (whether benign or suspicious) could confirm only that the probability of cancer was low. At MSK, the pre-test probability of an ITN being cancer was higher, leading to the inability of a GEC-benign result to rule out cancer with sufficient NPV (Fig. 2). The efficacy of the GEC is therefore variable from one institution to another, in contrast to Alexander et al.’s initial

report. These findings have significant implications for different practice settings. That is, the test in isolation may not provide sufficient information to alter management in institutions where the prevalence of malignancy in ITNs differs from the initially described validation cohort.

Molecular testing has gained acceptance as physicians seek better characterization of the escalating number of thyroid nodules diagnosed, with the goal of avoiding unnecessary interventions. Such molecular testing includes assays with a high PPV (to "rule in" cancer in ITNs), such as ThyGenX (Interpace Diagnostics, Parsippany, NJ), formerly known as the Asuragen miRinform panel. This assay evaluates the presence of *BRAF* and *RAS* mutations and rearrangements of RET/PTC and PAX8/PPARγ. The Veracyte Afirma GEC test is marketed as a "rule out" test for cancer in ITNs. It applies a benign gene expression profile using a panel of 167 genes assessed in mRNA isolated from a fine-needle aspirate. ThyroSeq v.2 (CBLPath, Ocala, FL) uses next-generation sequencing to test for characteristic point mutations and genetic fusions. This test appears to offer high NPVs and PPVs for thyroid cancer.

With any diagnostic assay, the sensitivity and specificity are characteristics intrinsic to the test. The "real world" performance of the test, measured in terms of PPV and NPV, depends heavily on the prevalence of disease in the population under study. Early data from several institutions has shown variable performance of the Afirma GEC.[7–] Some have suggested that this variability could indicate that the sensitivity and specificity of the GEC may be different than previously reported. Our data, however, showed that the PPV and NPV obtained match very closely with the predicted PPV and NPV, given the known prevalence of malignancy at each institution and the reported sensitivity and specificity of the test. Thus, we believe that clinicians can predict GEC efficacy at their institution based on the prevalence of malignancy in thyroid nodules at that institution.

Importantly, the efficacy of the Afirma GEC relies on its performance across different practice settings (Fig. 1). First, the ability of the test to potentially avoid surgery (with a GEC-benign result) is based on the benign call rate. This metric steadily declines from 50 to 25 % as the prevalence of malignancy in ITNs increases from 0 to 60 %. Second, the test's efficacy as a "rule out" test is based on its NPV. If an NPV >95 % is desired, it can best be anticipated in practice settings where the prevalence of malignancy in ITNs is <21 %. Third, in practice settings where the prevalence of malignancy is low, the post-test probability of malignancy (with a GEC-suspicious result) remains low (Fig. 3). This means that the test provides no additional information and merely confirms that the risk of cancer is low. For example, if the prevalence of malignancy at an institution is 10 %, the predicted PPV would only be 17 %. If a PPV of >25 % were desired, it could be expected in a practice setting where the prevalence of malignancy is >15 %.

Therefore, the Afirma GEC would be expected to provide the most useful information in a practice setting with a prevalence of malignancy in ITNs of 15–21 %. In this scenario, the performance characteristics would be predicted to approximate those reported by Alexander et al. The test may still provide some useful information in settings where the prevalence of malignancy is at 12–25 %. Outside this range, however, the test seems unlikely to provide

information that would alter management. In populations with a pre-test probability of <12 %, the PPV would be predicted to be <20 %, and therefore the risk of cancer remains low regardless of the GEC result. In populations with pre-test probability >25 %, the NPV is predicted to be <94 %, and so the test would fail to achieve the 95 % standard for a "rule out" test described in the National Comprehensive Cancer Network guidelines. These numbers are only predictors of performance. The actual efficacy of the GEC may vary. These calculations do, however, explain why the Afirma GEC did not provide information sufficiently effective to alter management at either MSK or MSBI. The prevalence of malignancy at either institution was not within the ideal range.

There are several caveats to this analysis. We do not know the true status of the majority of GEC-benign nodules that did not undergo surgery for confirmation, so we could only estimate the NPV. We based our estimates of test performance (Fig. 1) on the sensitivity and specificity of the Afirma GEC as reported by Alexander et al. These figures may vary somewhat across practice settings, perhaps based on institutional differences in cytopathologic classification of ITNs. We did find, however, that our observed values correlated well with predicted values at both MSK and MSBI, indicating that the sensitivity and specificity are likely to be close to the values reported and that variations in performance are less likely to be attributable to other sources such as differences in cytopathology. Both MSK and MSBI have high-volume, dedicated thyroid cytopathology programs, although some inter-institutional variation in thyroid cytopathology interpretation is inevitable. Outside of our two institutions, there may be wider variation in cytopathology and therefore wider variation in GEC performance. Indeed, this study probably underestimates the range of GEC performance in the community. With different practice patterns, referral patterns, prevalences of malignancy, and cytopathological variations, there may be even greater variance in observed GEC performance characteristics and efficacy.

CONCLUSIONS

Knowledge of the prevalence of malignancy in ITNs at a particular institution is critical for reliable interpretation of GEC results. The variation seen at the two institutions in our study suggest that GEC performance is not necessarily reproducible in all settings. Our data suggest, however, that the efficacy of the GEC assay can be predicted based on knowing the prevalence of malignancy in ITNs. We conclude that molecular data should not be used in isolation to drive clinical decision-making. It is best integrated with patient, radiological, and cytopathological factors to best guide management.

REFERENCES

1. Cibas ES, Ali SZ. The Bethesda system for reporting thyroid cytopathology. Thyroid. 2009; 19:1159–1165. [PubMed: 19888858]
2. Cibas ES, Ali SZ. NCI Thyroid FNA State of the Science Conference. The bethesda system for reporting thyroid cytopathology. Am J Clin Pathol. 2009; 132:658–665. [PubMed: 19846805]
3. Alexander EK, Kennedy GC, Balock ZW, et al. Preoperative diagnosis of benign thyroid nodules with indeterminate cytology. N Engl J. Med. 2012; 367:705–715. [PubMed: 22731672]
4. McIver B. Evaluation of the thyroid nodule. Oral Oncol. 2013; 49:645–653. [PubMed: 23706806]

5. Ho AS, Sarti EE, Jain SK, et al. Malignancy rate in thyroid nodules classified as Bethesda category III (AUS/FLUS). Thyroid. 2014; 24:832–839. [PubMed: 24341462]
6. Iskander M, Bonomo G, Avadhani V, Persky M, Lucido D, Wang B, Marti JL. Evidence for the overestimation of malignancy in indeterminate thyroid nodules classified as bethesda III. Surgery. 2015; 157:510–517. [PubMed: 25633738]
7. Sox HC Jr. Probability theory in the use of diagnostic tests: an introduction to critical study of the literature. Ann Intern. Med. 1986; 104:60–66. [PubMed: 3079637]
8. Nikiforov YE, Ohori NP, Hodak SP, et al. Impact of mutational testing on the diagnosis and management of patients with cytologically indeterminate thyroid nodules: a prospective analysis of 1056 FNA samples. J Clin Endocrinol Metab. 2011; 96:3390–3397. [PubMed: 21880806]
9. Nikiforov YE, Carty SE, Chiosea SI, et al. Highly accurate diagnosis of cancer in thyroid nodules with follicular neoplasm/suspicious for a follicular neoplasm cytology by ThyroSeq v2 next-generation sequencing assay. Cancer. 2014; 120:3627–3634. [PubMed: 25209362]
10. Alexander EK, Schorr M, Klopper J, et al. Multicenter clinical experience with the Afirma gene expression classifier. J Clin Endocrinol Metab. 2014; 99:119–125. [PubMed: 24152684]
11. McIver B, Castro MR, Morris JC, et al. An independent study of a gene expression classifier (Afirma) in the evaluation of cytologically indeterminate thyroid nodules. J Clin Endocrinol Metab. 2014; 99:4069–4077. [PubMed: 24780044]
12. Lastra RR, Pramick MR, Crammer CJ, LiVolsi VA, Baloch ZW. Implications of a suspicious afirma test result in thyroid fine-needle aspiration cytology: an institutional experience. Cancer Cytopathol. 2014; 122:737–744. [PubMed: 25123499]
13. Harrell RM, Bimston DN. Surgical utility of Afirma: effects of high cancer prevalence and oncocytic cell types in patients with indeterminate thyroid cytology. Endocr Pract. 2014; 20:364–369. [PubMed: 24246351]
14. National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology: Thyroid Cancer. Version 2.2014. Available: www.nccn.org/professionals/physician_gls/f_guidelines.asp.

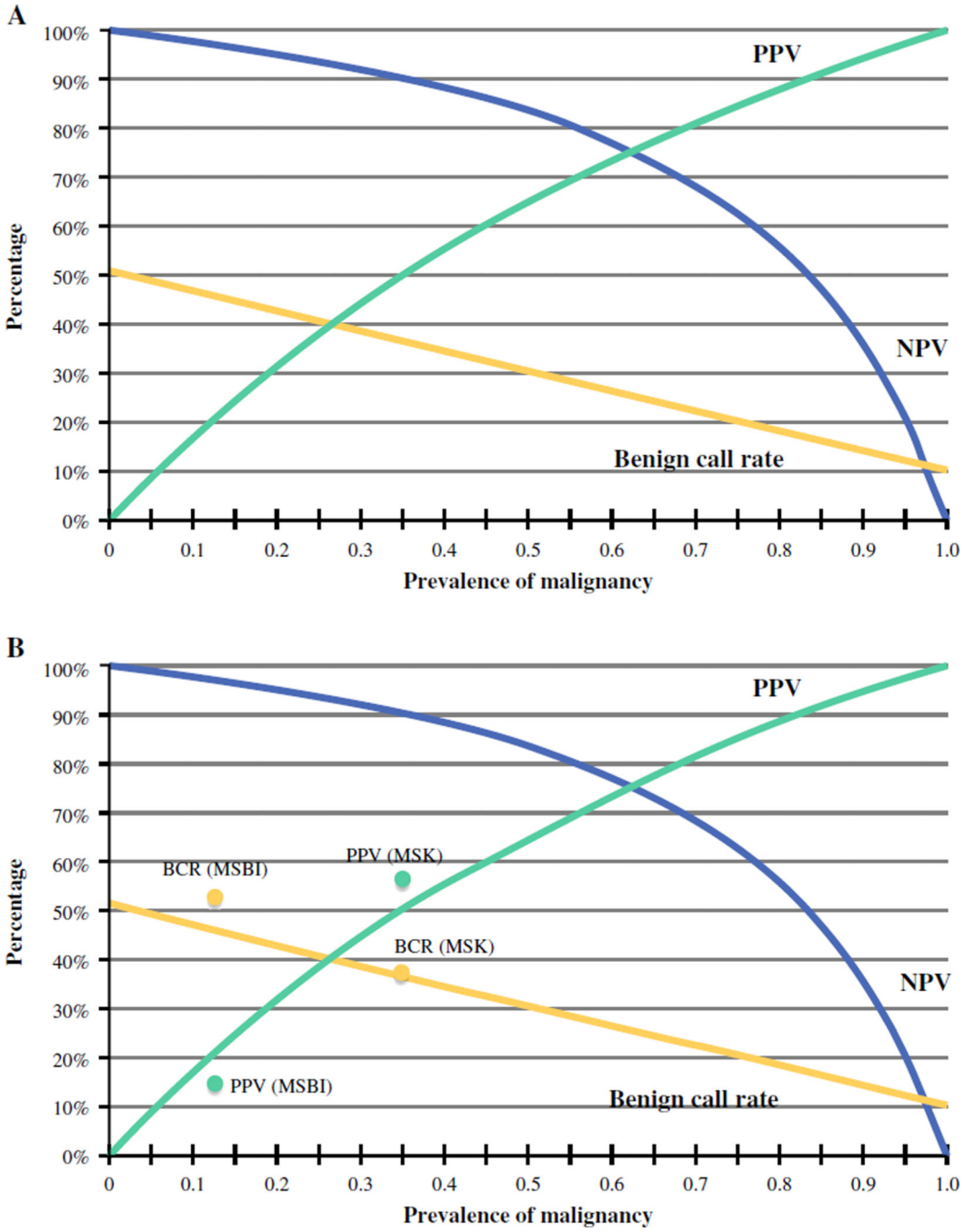

FIG. 1.

a Estimated gene expression classifier (GEC) performance characteristics based on the prevalence of malignancy, with GEC sensitivity of 90 % and specificity 51 %. *PPV* positive predictive value; *NPV* negative predictive value, *BCR* benign call rate (percentage of indeterminate nodules that are GEC-benign). **b** Actual benign call rate (BCR) and PPV values at Memorial Sloan Kettering Cancer Center (MSK) and Mount Sinai Beth Israel (MSBI)

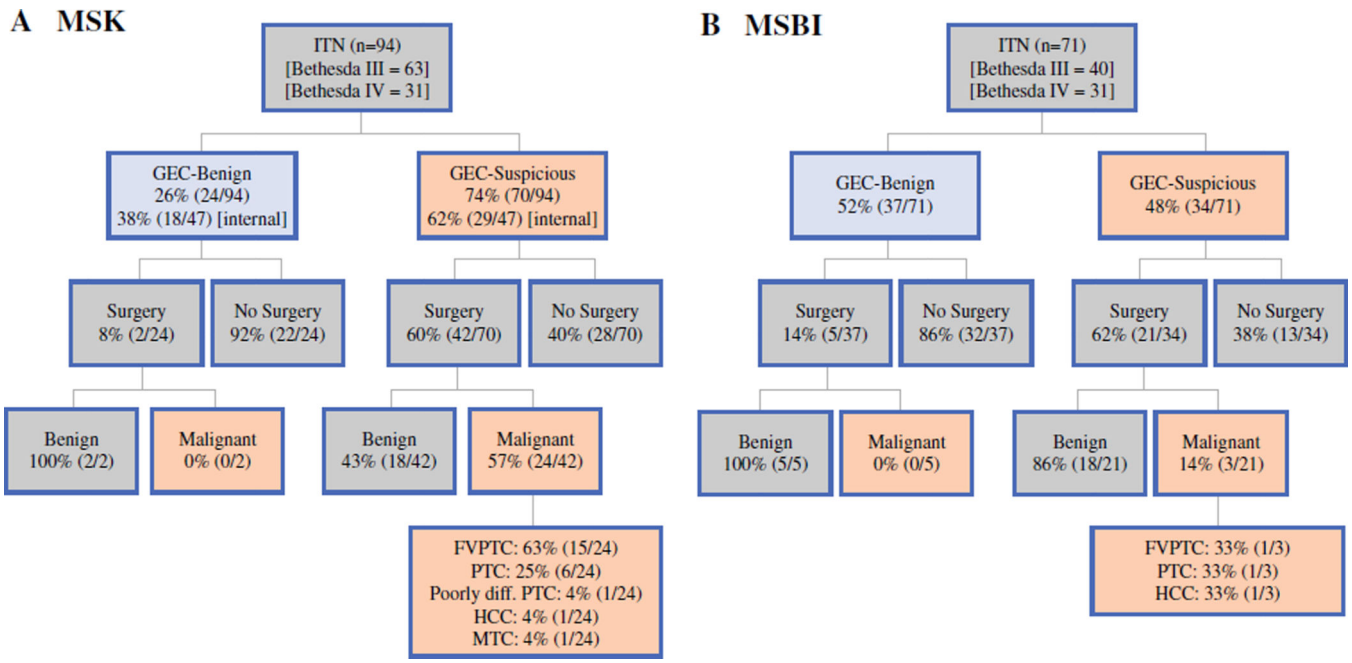

FIG. 2.

Management of indeterminate thyroid nodules at (**a**) MSK and (**b**) MSBI. *ITN* indeterminate thyroid nodules, *GEC* gene expression classifier, *internal* patients undergoing workup and Afirma evaluation at MSK or MSBI, in contrast to patients referred already with workup and Afirma evaluation completed, *FVPTC* follicular variant of papillary thyroid carcinoma, *PTC* papillary thyroid carcinoma, *Poorly diff.* poorly differentiated, *HCC* Hürthle cell carcinoma, *MTC* medullary thyroid carcinoma

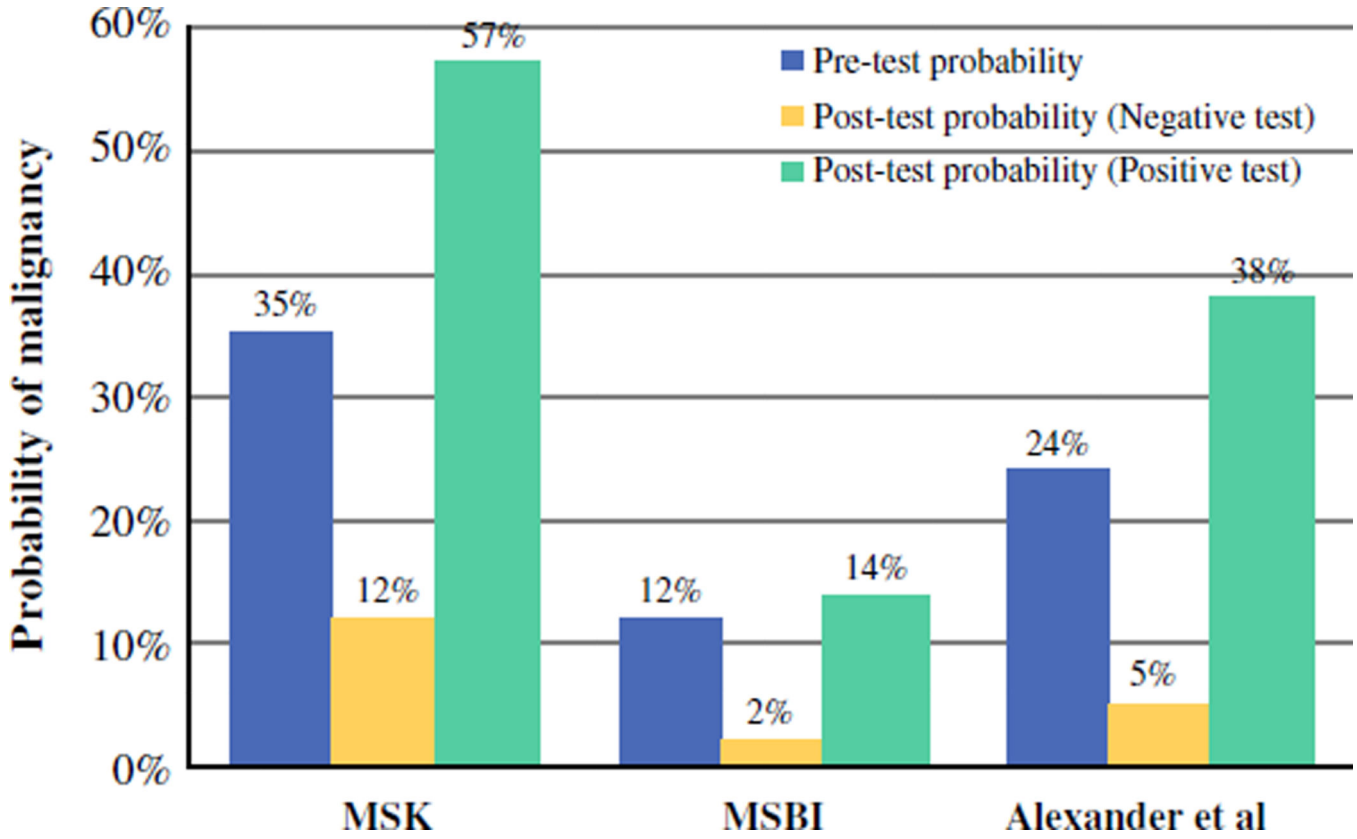

FIG. 3.

Comparative pre-test and post-test probability of malignancy. Pre-test probabilities of malignancy at MSK and MSBI were substantially different from those of Alexander et al., leading to wide variation in GEC clinical performance. Post-test probability (negative test) represents the probability of malignancy given a GEC-benign result, or 1–NPV. Post-test probability (positive test) represents the probability of malignancy given a GEC-suspicious result, or PPV

TABLE 1

Comparison of MSK and MSBI demographic and performance characteristics

Characteristic	MSK	MSBI
No. of patients	94	62
Female (%)	66	72
Mean age (years)	49	56
No. of nodules	94	71
Mean nodule size (cm)	2.1	2.3
% GEC-benign nodules that were malignant	0	0
% GEC-suspicious nodules that underwent surgery	60.0	61.7

MSK Memorial Sloan Kettering Cancer Center, *MSBI* Mount Sinai Beth Israel, *GEC* gene expression classifier