THYROID
Volume 32, Number 10, 2022

DOI: 10.1089/thy.2022.0269

Assessing Bias and Limitations of Clinical Validation Studies of Molecular Diagnostic Tests for Indeterminate Thyroid Nodules: Systematic Review and Meta-Analysis

Catherine DiGennaro,[1,2] Vahab Vahdatzad,[1,2] Mohammad S. Jalali,[1,3] Asmae Toumi,[1,2] Tina Watson,[1,2] G. Scott Gazelle,[1,3] Nathaniel Mercaldo,[1,3] and Carrie Cunningham Lubitz[1,2]

Background: Molecular tests for thyroid nodules with indeterminate fine needle aspiration results are increasingly used in clinical practice; however, true diagnostic summaries of these tests are unknown. A systematic review and meta-analysis were completed to (1) evaluate the accuracy of commercially available molecular tests for malignancy in indeterminate thyroid nodules and (2) quantify biases and limitations in studies that validate those tests.

Summary: PubMed, EMBASE, and Web of Science were systematically searched through July 2021. English language articles that reported original clinical validation attempts of molecular tests for indeterminate thyroid nodules were included if they reported counts of true-negative, true-positive, false-negative, and false-positive results. We performed screening and full-text review, followed by assessment of eight common biases and limitations, extraction of diagnostic and histopathological information, and meta-analysis of clinical validity using a bivariate linear mixed-effects model. Forty-nine studies were included. Meta-analysis of Afirma Gene expression classifiers (GEC; $n = 38$ studies) revealed a sensitivity of 0.92 (confidence interval: 0.90–0.94), specificity of 0.26 (0.20–0.32), negative likelihood ratio (LR−) of 0.32 (0.23–0.44), positive LR+ of 1.24 (1.15–1.35), and area under the curve (AUC) of 0.83 (0.74–0.89). Afirma Genomic Sequencing Classifier (GSC; $n = 10$) had a sensitivity of 0.94 (0.89–0.96), specificity of 0.38 (0.27–0.50), LR− of 0.18 (0.10–0.30), LR+ of 1.52 (1.28–1.87), and AUC of 0.91 (0.62–0.92). ThyroSeq v1 and v2 ($n = 10$) had a sensitivity of 0.86 (0.82–0.90), specificity of 0.74 (0.59–0.85), LR− of 0.19 (0.13–0.26), LR+ of 3.52 (2.08–5.92), and AUC of 0.86 (0.81–0.90). ThyroSeq v3 ($n = 6$) had a sensitivity of 0.92 (0.86–0.95), specificity of 0.41 (0.18–0.69), LR− of 0.24 (0.09–0.62), LR+ of 1.67 (1.09–2.98), and AUC of 0.90 (0.63–0.92). Fourteen percent of studies conducted a blinded histopathologic review of excised thyroid nodules, and 8% made the decision to go to surgery blind to molecular test results.

Conclusions: Meta-analyses reveal a high diagnostic accuracy of molecular tests for thyroid nodule assessment of malignancy risk; however, these studies are subject to several limitations. Limitations and their potential clinical impacts must be addressed and, when feasible, adjusted for using valid statistical methodologies.

Keywords: bias assessment, clinical validity, diagnostic tests, meta-analysis, molecular diagnostics, oncology, systematic review, thyroid cancer

[1]Institute for Technology Assessment, Massachusetts General Hospital, Boston, Massachusetts, USA.
Departments of [2]Surgery and [3]Radiology, Massachusetts General Hospital, Boston, Massachusetts, USA.

Introduction

DIAGNOSTIC TESTS DETECT and monitor disease and are ubiquitous across fields of medicine. In 2017, molecular diagnostic tests' $3.7b market share in North America was composed primarily of infectious disease and genetic testing, and the development of advanced cancer diagnostic tests is projected to increase more drastically than any other segment of the market.[1] While diagnostic tests have shown benefits by reducing unnecessary surgeries and hospitalizations[2–4] improving the accuracy of decisions made by practitioners,[5,6] they may not always be a helpful component of treatment. Indeed, diagnostic tests can be over-ordered, used inappropriately or without knowledge of limitations, and may be associated with medical risks and high costs that undermine their utility in guiding medical decision-making.[7]

Sensitivity and specificity are key components of a diagnostic test with one prioritized over the other based on the clinical scenario. Diagnostic tests aim to minimize risk and cost, which are different in a false-negative diagnosis versus a false-positive diagnosis; one measure of validity can be prioritized over another to increase the test's utility, rather than accuracy.[8] Clinically, the impact of each test must be considered and can be further evaluated by calculating the negative likelihood ratio (LR−), for example, to evaluate the risk of a false-negative result. For instance, a false positive may result in unnecessary treatment; a false negative may result in a missed cancer.

True test performance may be masked by biases within the initial clinical utility validation, in addition to changes in underlying prevalence. In the absence of a prospective, double-blinded study design, prospective–retrospective studies with banked biospecimens, single-arm studies, prospective observational studies, or decision-analytic modeling techniques[9] may be conducted, increasing the associated risks of biases. As a result, the actual sensitivity, specificity, or area under the curve (AUC) of diagnostic tests may differ in truth from reported values, resulting in unnecessary treatment and health care utilization of patients without an actual underlying disease and, for misdiagnosed patients with underlying disease, delayed treatment and reduced treatment efficacy.

Accounting for the consequences of a false-negative test result (i.e., a missed cancer), molecular tests (MTs) for thyroid cancer were developed with a high sensitivity to rule out malignancy by assessing the presence of biomarkers and genetic mutations in nodules with indeterminate fine needle aspiration (FNA) results.[10] Bethesda III and IV, in the six-tiered Bethesda System for Reporting Thyroid Cytology, are considered indeterminate FNA results.[11] The American Thyroid Association recommended the use of molecular tests for indeterminate thyroid nodules in clinical practice in 2015.[12] Gene expression classifiers (GEC) were designed to have high sensitivity and a negative predictive value, thus ruling out malignancy and presumed avoidance of unnecessary surgery.[13] Veracyte developed the first widely used molecular test, the Afirma GEC molecular test in 2012.[13] The test was widely adopted after successful clinical trials. Other companies soon followed suit.[14]

To determine which biases are present in the reports of diagnostic tests and to what degree, careful reviews of test protocols should be conducted. It is critical that the reported accuracy of these tests be interpreted within the context of potential biases. Previous reviews[15–17] are yet to comprehensively and systematically identify and analyze these biases across all the diagnostic molecular tests for thyroid carcinoma.

In this study, we characterize and quantify limitations and biases within the studies used to validate molecular tests that evaluate cytologically indeterminate thyroid nodules.

Materials and Methods

This review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[18] This review was not preregistered and an associated protocol is not available.

Data sources and searches

A systematic literature search was conducted on PubMed, Embase, and Web of Science database by searching for articles reporting attempts to clinically validate commercially available molecular tests for thyroid nodules. The search strategy consisted of querying each database for combinations of the keywords ''molecular testing,'' ''cytomolecular testing,'' ''pathology,'' ''genetic testing,'' or ''genetic expression''; AND ''indeterminate,'' ''cytology,'' ''fine needle aspiration,'' ''thyroid nodule,'' or ''thyroid surgery''; AND ''Bethesda category III,'' ''Bethesda category IV,'' ''benign thyroid nodule,'' ''papillary thyroid cancer,'' ''nonmedullary thyroid carcinoma,'' or ''thyroid cancer.'' Full details of the search queries are presented in Supplementary Appendix Methods SAD. This allowed for a broad sample of articles that contained clinical validation results of molecular tests for indeterminate thyroid nodules.

Date limits were not used, and the search was finalized on July 29, 2021. Reference lists of included articles were examined to ensure the inclusion of all relevant articles.

Study selection

Original articles reporting clinical validity studies of commercially available molecular tests were selected for inclusion if they reported diagnostic results, that is, counts of true-negative (TN), true-positive (TP), false-negative (FN), and false-positive (FP) results. Clinical validity studies assess a diagnostic test's ability to distinguish between patients with and without a disease,[19] and the process of performing one such study is characterized elsewhere.[20] Commercially available molecular tests identified, which are either currently or have been previously used, include Afirma GEC, Afirma Genomic Sequencing Classifier (GSC), ThyroSeq versions 1–3, miRInform or ThyGenX Thyroid Oncogene Panel (miRInform), Quest Diagnostics Genetic Mutation Panel, Rosetta GX Reveal, and ThyraMIR. Articles written in a language other than English and conference abstracts were excluded.

Two researchers (C.D. and V.V.) independently assessed the titles, abstracts of a pilot sample of articles for inclusion, and full-text articles. In cases where there was a discrepancy in their inclusion assessments, a third researcher (C.C.L.) resolved conflicts. Articles with patient samples that overlap with another study validating the same test were excluded; in each case of overlapping patient samples, the study with a larger sample was retained.

Data extraction and quality assessment

For each included study, the following information was extracted independently by one researcher (C.D.): authors and year of publication; true-positive, false-positive, true-negative, and false-negative results; sensitivity and specificity; positive and LR−; positive and negative predictive value; patient and nodule sample size; nodule cytological result; institutional prevalence of malignancy, initial FNA results, and type of study (i.e., prospective or retrospective); commercial molecular test(s), country-level location of study, and number of institutions; noninvasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP) classification (only assessed in articles published after 2015); industry funding receipt; and molecular test results and final histopathology diagnoses.

Risk of bias of the included studies was assessed by one researcher (C.D.) using nine categorical criteria covering the following aspects: patient selection, inconsistent comparison bias, partial verification bias, diagnostic review bias, observer variability, reporting of indeterminate results, and institutional malignancy prevalence reported as a range (Supplementary Appendix Table SA1). Each study's design and execution were assessed, and the risk of each bias was recorded as a binary variable (Met or Not Met) for whether each criterion was met with the option to partially meet a criterion, for example, in the case of pooled results from multiple locations with different study protocols.

Articles were given a summary score based on the number of criteria they met out of nine (if the article specified that multiple histopathologists reviewed excised samples) or eight (if only one histopathologist reviewed excised samples). Partial verification bias was assessed by quantifying the proportion of molecular test results verified by a final histopathological diagnosis.

Types of biases and limitations

Several biases can occur before, during, and after conducting diagnostic tests and, depending on the initial study design, can be difficult to avoid. The focus of this study is to characterize these biases. We distinguish between ''biases,'' ''limitations,'' and ''variations''; bias arises from defects in study design and can result in incorrect conclusions; the direction of bias can sometimes be identified[21] and statistically corrected.[20,22–26] Limitations arise from oversights in study design and cannot typically be corrected for, and variation is a by-product of changing underlying conditions among studies, for example, patient population and clinical protocol. These underlying conditions should be reported to contextualize findings.[27] Primary biases and limitations in study design are detailed in Table 1.

Data synthesis and analysis

Statistical analyses were performed using R version 3.6.2.[28] Publication bias was assessed separately for each set of studies evaluating the same molecular test using Deek's funnel plot, which displays the inverse of the square root of sample size (number of nodules) as a function of the diagnostic odds ratio (the odds of a positive test in patients with disease relative to the odds of a negative test in patients without disease).[29] Cochran's Q test[30] was used to assess underlying population heterogeneity between studies evaluating the same molecular test. For each molecular test with more than three studies assessing its clinical validity, true-positive, false-positive, true-negative, and false-negative results were extracted (results were assessed per nodule, even if there were multiple nodules tested from a single patient). Studies without information needed to calculate both sensitivity and specificity were excluded from the analysis. Exact confidence intervals (CIs) were computed for sensitivity and specificity.[31,32]

A bivariate linear mixed-effects model was constructed to jointly model the logit-transformed sensitivity, specificity, and positive and negative LRs across studies.[33] AUC was extracted, and a CI was constructed using a bootstrapping approach,[34] described previously.[35] The clinical validity results of molecular tests with fewer than three validation studies are reported as extracted and were not summarized.

Results

Study selection and inclusion

The literature search yielded 2062 results. After excluding 532 duplicates, the abstracts of 1530 unique records were screened for eligibility, which resulted in 103 articles included for full-text review. The final screening resulted in 49 articles (Fig. 1). See Supplementary Appendix Table SA1 for a list of excluded full-text articles and reasons for their exclusion.

Study characteristics are described in Table 2. Thirty-five studies[13,36–69] evaluated the clinical validity of Afirma GEC; nine[38,44 47,56,63 64,69–71] evaluated Afirma GSC. One study[72] evaluated both Afirma GEC and GSC, noting an institutional switch in clinical practice without separating the results, so it is included in the analyses of Afirma GEC and Afirma GSC. Nine studies[49,50,54,73–78] evaluated ThyroSeq v1 or v2 (version 1 and version 2 results were combined in the studies that did evaluate both[49,77,79]), and five studies[64,71,78,80,81] evaluated ThyroSeq v3 (a significant expansion of previous versions). One study[82] evaluated both ThyroSeq v2 and ThyroSeq v3, noting an institutional switch in clinical practice without separating the results, so it is included in the analyses of ThyroSeq v1 and v2 and ThyroSeq v3. Two[58,72] evaluated Rosetta GX Reveal, two[83,84] evaluated miRInform, one[72] evaluated ThyraMIR, and one[73] evaluated Quest Diagnostics' gene mutation panel (GMP).

Of the 49 studies included, 14[38,44,47,49–51,54,56,58,63,69,71,73,78] compared 2 diagnostic tests and 2[64,72] compared 3 diagnostic tests. Thirty-nine studies[36–38,40–45,47–49,51,53–59,61–70,72–74,77,78,81–84] used a retrospective design, nine studies[13,39,46,50,52,60,71,75,80] used a prospective design, and one study[76] combined retrospective and prospective cohorts. Twenty-five studies[13,36,37,41,42,44–47,50,52,56,61,63,64,66,68–71,75,77,78,80,82] performed a single FNA before molecular testing and two[13,70] reported receiving industry sponsorship. The proportion of quality criteria met varies greatly among these studies (Table 2).

Quality assessment

Risk of bias was quantified using categorical criteria as outlined in Supplementary Appendix Table SA2. The risk-of-bias assessment results are displayed in Figure 2, and detailed results can be found in Supplementary Appendix Table SA3

Table 1. Biases and Limitations in Study Design Assessed

	Description	*Impact on accuracy*	*Preventative measures*	*Impact on thyroid cancer diagnosis*
Biases				
Selection	Eligible participants are not enrolled randomly or consecutively	Depends on who is not represented	Ensure that participants are enrolled randomly or consecutively, report participant exclusion criteria	If the study population does not represent the underlying characteristics of the available population, the results may not be generalizable to another population.
Verification	A nonrandom set of patients do not undergo the reference test[a]	Overestimation of sensitivity; effect on specificity can vary	Perform the reference test on all patients (unnecessary, invasive, and costly procedures are unethical in clinical practice); or do not base the decision to perform the reference test on the results of the index test	Patients who receive a benign test result rarely go to surgery, while patients who receive a malignant test result typically go to surgery; since the gold standard reference test is invasive, it is not applied to everyone. Thus, identifying a false-negative result is far less likely than identifying a false-positive result.
Observer variability	Multiple observers separately interpret reference test results and do not compare findings	Depends on the direction/degree of differential classification	Compare interpretations of the reference tests and ensure that the standards are consistent; alternatively, perform a sensitivity analysis to quantify the effect of observer variability to assess the overall impact on results	When multiple pathologists interpret histopathologic results, there may be differences in their conclusions; this can bias the assessed accuracy of the molecular test.
Indeterminate results excluded	Nondiagnostic index test results are excluded from the analysis	Overestimation of accuracy reduces generalizability of results	Report and explain indeterminate test results	Molecular tests for thyroid cancer return nondiagnostic results frequently; to determine the clinical *utility* of a molecular test, nondiagnostic results must be reported.
Limitations				
Diagnostic review	Interpretation of reference test is not blind to index test results	Overestimation of accuracy	Interpretation of reference test should be blind to index test results	Histopathologic diagnoses of thyroid nodules are interpreted after molecular tests in prospective studies and retrospective studies where the test was conducted during the clinical time line and may be interpreted with knowledge of the molecular test result. This can bias histopathological interpretation toward the molecular test result, thus overestimating the test's accuracy.
Context	Institutional prevalence is not reported as a range	Depending on whether the prevalence is lower or higher than average, this may limit generalizability	Report institutional prevalence as a range that reflects the institutional context	Measures such as positive and negative predictive values depend on institutional prevalence; the exact institutional prevalence of thyroid cancer is often unknown. Reporting upper and lower confidence intervals promotes decision-making over a range and correctly illustrates uncertainty inherent to sampled measures of prevalence.
Inconsistent comparison	Results from different index tests[b] are combined into one summary	Depends on the accuracy of the underlying tests	Ensure that comparisons are made between results of the same test, or else separated by test	Multiple types of molecular tests are used within a single institution; combining them in analysis is inappropriate and can lead to incorrect assessments of a single test's utility.

[a]Histopathologic diagnosis (generally, the "gold standard" test revealing ground truth).
[b]Molecular diagnostic test (generally, the test being evaluated).
Adapted from Campbell et al[91] and Santaguida et al.[92]

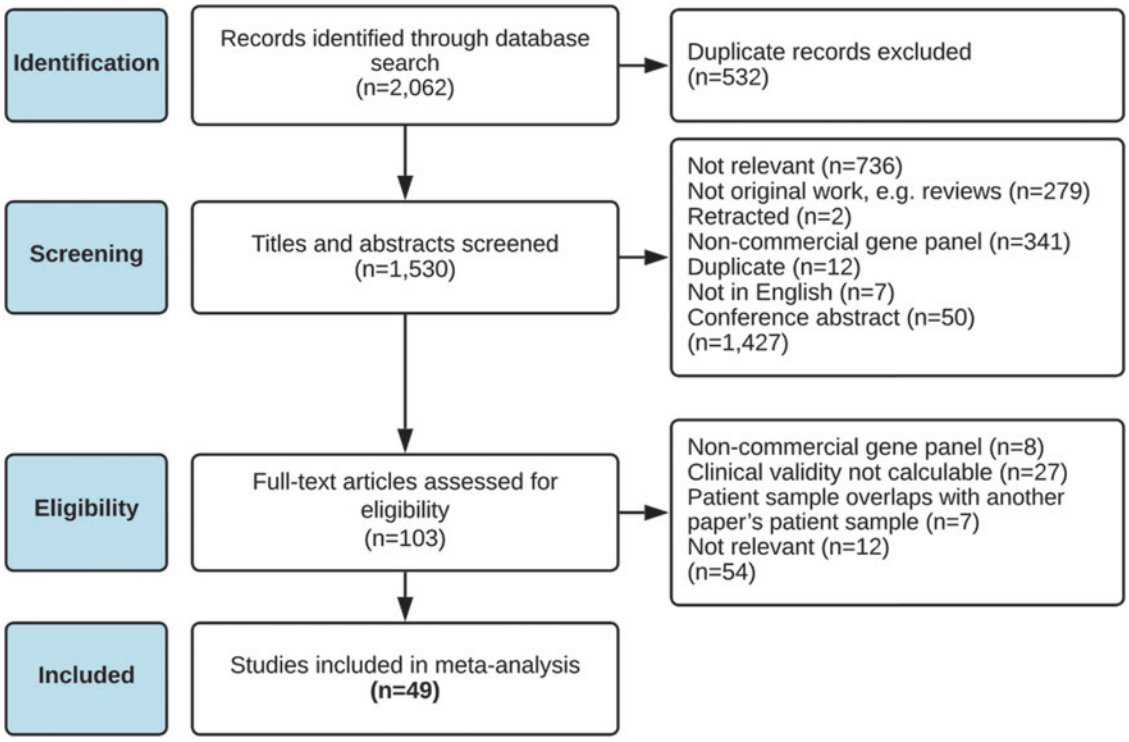

FIG. 1. PRISMA flow diagram. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

and Supplementary Appendix Figure SA1. Of the 49 studies evaluated, all of the studies enrolled patients consecutively, and 90% of studies evaluated thyroid nodules with a consistent molecular test, or else reported the results of different molecular tests separately. Eighty-four percent enrolled patients regardless of whether they ultimately received surgery (rather than only including patients who ultimately did receive surgery). Seventy-one percent of studies reported their exclusion criteria. Thirty-eight percent of studies that reported multiple histopathologists or multiple institutions ($n=34$) addressed observer variability in their histopathology review. Forty-one percent of studies reported nondiagnostic molecular test results, and 31% of studies reported the institutional prevalence of malignancy (i.e., pretest probability of cancer) as a range to reflect uncertainty.

Fourteen percent of the studies conducted histopathologic reviews blind to the molecular test results, and 8% made the decision to go to surgery blind to the molecular test result (Fig. 2). Six percent of studies fulfilled more than 75% of the quality assessment criteria, and 49% of the studies fulfilled more than 50% (Table 2).

Study population characteristics and molecular test results are presented in Supplementary Appendix Tables SA4 and SA5. The institutional context varies greatly across studies; underlying prevalence of malignancy ranges from 4.7% to 67% (mean, 0.21), and patients are sent to surgery between 7% and 96% of the time (excluding studies that only included patients who had surgery). Furthermore, the initial nodule cytology breakdown varies across institutions, as does the study reporting; while most studies only tested cytologically indeterminate (Bethesda III and IV) nodules, some studies did perform molecular testing on Bethesda I, II, V, and VI nodules (Supplementary Appendix Table SA4).

Statistical analysis

Meta-analysis results are presented in Figure 3. Cochran's Q statistic revealed significant heterogeneity between studies evaluating Afirma GEC ($n=38$); $p=0.006$ for sensitivity and $p<0.001$ for specificity. A random-effects bivariate model revealed a summary sensitivity of 0.92 (CI: 0.90–0.94), a specificity of 0.26 (CI: 0.20–0.32) (Fig. 3A), a negative LR– of 0.32 (0.23–0.44), a positive LR+ of 1.24 (1.15–1.35), and an AUC of 0.83 (0.74–0.89). We were unable to detect heterogeneity between studies evaluating Afirma GSC ($n=10$; $p=0.88$), but there was significant heterogeneity between specificities ($p<0.001$). The sensitivity of Afirma GSC was 0.94 (0.89–0.96), the specificity was 0.38 (0.27–0.50) (Fig. 3B), the LR– was 0.18 (0.10–0.30), the LR+ was 1.52 (1.28–1.87), and the AUC was 0.91 (0.62–0.92).

We were unable to detect heterogeneity between sensitivities of studies evaluating ThyroSeq v1 and v2 ($n=10$; $p=0.98$), but significant heterogeneity was detected between specificities ($p<0.001$). The overall sensitivity of ThyroSeq v1 and v2 was 0.86 (0.82–0.90), and the overall specificity was 0.74 (0.59–0.85) (Fig. 3C). The LR– was 0.19 (0.13–0.26), the LR+ was 3.52 (2.08–5.92), and the AUC was 0.86 (0.81–0.90). We were unable to detect heterogeneity between sensitivities of studies evaluating ThyroSeq v3 ($n=6$; $p=0.54$), but significant heterogeneity was detected between specificities ($p<0.001$). The overall sensitivity of ThyroSeq v3 was 0.92 (0.86–0.95), and the overall specificity was 0.41 (0.18–0.69) (Fig. 3D). The LR– was 0.24 (0.09–0.62) and the LR+ was 1.67 (1.09–2.98), while the AUC was 0.90 (0.63–0.92).

Full results of all four models can be found in Supplementary Appendix Tables SA6–SA9. Results pooled by

Table 2. Study Characteristics

Authors (year)[Ref.]	*Country*	*Sites*	*Study design*	*Molecular test*	*Patients*	*Nodules*	*FNAs before MT*	*NIFTP designation*	*Industry sponsorship*	*Quality criteria met (%)*[a,b]
Al-Qurayshi et al (2017)[36]	United States	1	Retrospective	AGEC	145	154	1	—	No	50
Alexander et al (2012)[13]	United States	49	Prospective	AGEC	249	265	1	Pre-2016	Yes	78
Alexander et al (2014)[37]	United States	5	Retrospective	AGEC	339	339	1	Pre-2016	No	56
Angell et al (2019)[38]	United States	1	Retrospective	AGEC	563	486	>1	Malignant	No	63
				AGSC		114				
Arosemena et al (2020)[62]	United States	1	Retrospective	AGEC	117	126	Unreported	Malignant	No	44
Azizi et al (2018)[39]	United States	1	Prospective	AGEC	151	151	>1	Benign	No	44
Baca et al (2017)[40]	United States	1	Retrospective	AGEC	219	227	>1	Malignant	No	50
Brauner et al (2015)[41]	United States	3	Retrospective	AGEC	72	72	1	Pre-2016	No	67
Carty et al (2020)[78]	United States	1	Retrospective	TSv2	Unreported	240	1	Both	No	56
				TSv3	Unreported	119				
Celik et al (2015)[42]	United States	1	Retrospective	AGEC	66	66	1	Pre-2016	No	50
Chaudhary et al (2016)[43]	United States	1	Retrospective	AGEC	Unreported	158	>1	Malignant	No	56
Chen et al (2020)[81]	Canada	1	Retrospective	TSv3	50	50	>1	Unclear	No	56
Endo et al (2019)[44]	United States	1	Retrospective	AGEC	317	367	1	Malignant	No	44
				AGSC	153	165				
Geng et al (2021)[63]	United States	1	Retrospective	AGEC	167	167	1	Malignant	No	33
				AGSC	133	133				
Gortakowski et al (2021)[64]	United States	1	Retrospective	AGEC	89	92	1	Benign	No	33
				AGSC	70	73				
				TSv3	55	59				
Hang et al (2017)[45]	United States	1	Retrospective	AGEC	375	384	1	Both	No	67
Harrell et al (2014)[46]	United States	1	Prospective	AGEC	58	58	1	Pre-2016	No	56
Harrell et al (2019)[47]	United States	1	Retrospective	AGEC	Unreported	481	1	Malignant	No	56
				AGSC		139				
Harrison et al (2017)[48]	United States	1	Retrospective	AGEC	110	115	>1	—	No	44
Jug et al (2018)[49]	United States	1	Retrospective	AGEC	198	207	>1	Both	No	56
				TSv1 and v2	91	97				
Labourier et al (2015)[83]	United States	12	Retrospective	miRInform	Unreported	109	—	Pre-2016	No	67
Li et al (2021)[82]	United States	1	Retrospective	TSv2 and v3	Unreported	202	1	Benign	No	33
Livhits et al (2018)[50]	United States	1	Prospective	AGEC	70	70	1	Malignant	No	78
				TSv2	79	79				
Livhits et al (2021)[71]	United States	1	Prospective	AGSC	189	201	1	Benign	No	67
				TSv3	157	171				
Maerki et al (2019)[73]	United States	1	Retrospective	Quest GMP	86	86	—	Benign	No	33
				TSv2	79	79				
Marcadis et al (2019)[74]	United States	4	Retrospective	TSv2	266	273	—	Both	No	56
Marti et al (2015)[51,c]	United States	2	Retrospective	AGEC	94	94	—	Pre-2016	No	44
				AGEC	62	71				
McIver et al (2014)[52]	United States	1	Prospective	AGEC	72	72	1	Pre-2016	No	75
Nikiforov et al (2014)[76]	United States	1	Both	TSv2	143	143	—	Pre-2016	No	50
Nikiforov et al (2015)[75]	United States	1	Prospective	TSv2	441	465	1	Pre-2016	No	50
Nishino et al (2021)[65]	United States	1	Retrospective	AGEC	360	370	>1	Malignant	No	67

(*continued*)

TABLE 2. (CONTINUED)

Authors (year)[Ref.]	*Country*	*Sites*	*Study design*	*Molecular test*	*Patients*	*Nodules*	*FNAs before MT*	*NIFTP designation*	*Industry sponsorship*	*Quality criteria met (%)*[a,b]
Noureldine et al (2015)[53]	United States	1	Retrospective	AGEC	273	274	>1	Pre-2016	No	50
Papoian et al (2020)[66]	United States	1	Retrospective	AGEC	69	69	1	Unclear	No	44
Parajuli et al (2019)[54]	United States	1	Retrospective	AGEC	199	202	>1	Benign	No	50
				TSv1 and v2	81	81				
Partyka et al (2019)[72]	United States	1	Retrospective	AGEC and AGSC	68	68	—	Benign	No	44
				Reveal	23	23				
				ThyraMIR	22	22				
Patel et al (2018)[70]	United States	49	Retrospective	AGSC	183	190	1	Malignant	Yes	67
Samulski et al (2016)[55]	United States	1	Retrospective	AGEC	294	294	>1	Malignant	No	63
San Martin et al (2020)[56]	United States	1	Retrospective	AGEC	174	178	1	—	No	50
				AGSC	116	121				
Shrestha et al (2016)[77]	United States	1	Retrospective	TSv1 and v2	261	261	1	—	No	38
Steward et al (2019)[80]	United States, Singapore	10	Prospective	TSv3	257	286	1	Malignant	No	100
Sultan et al (2020)[67]	United States	1	Retrospective	AGEC	98	101	>1	Both	No	33
Valderrabano et al (2016)[84]	United States	1	Retrospective	miRInform	105	116	>1	—	No	75
Villabona et al (2016)[57]	United States	1	Retrospective	AGEC	48	48	1 and >1	—	No	38
Vora et al (2020)[68]	United States	1	Retrospective	AGEC	368	416	1	Unclear	No	56
Walts et al (2018)[58]	United States	3	Retrospective	AGEC	79[d]	81[d]	—	Both	No	33
				Reveal						
Witt (2016)[59]	United States	1	Retrospective	AGEC	32	32	>1	—	No	63
Wu et al (2016)[60]	United States	1	Prospective	AGEC	231	245	—	—	No	56
Yang et al (2016)[61]	United States	1	Retrospective	AGEC	217	217	1	—	No	67
Zhang et al (2021)[69]	United States	1	Retrospective	AGEC	120	127	1	Benign	No	44
				AGSC	125	137				

[a]Percent of criteria (defined in Supplementary Appendix Table SA1) fulfilled, does not include observer variability criterium for studies with only one histopathologist ($N = 15$).
[b]Full-quality assessment results, including whether criteria were met, not met, partially met, not reported, or not applicable can be found in Supplementary Appendix Table SA2.
[c]Results separated by institution.
[d]AGEC and Reveal applied to same nodules.

AGEC, Afirma Gene Expression Classifier; AGSC, Afirma Genomic Sequencing Classifier; FNA, fine needle aspiration; GMP, gene mutation panel; miRInform, miRInform or ThyGenX Thyroid Oncogene Panel; MT, molecular test; N, no; NIFTP, noninvasive follicular thyroid neoplasm with papillary-like nuclear features; Reveal, RosettaGx Reveal; TS, ThyroSeq (versions 1 and 2); TS v3, ThyroSeq version 3; Y, yes.

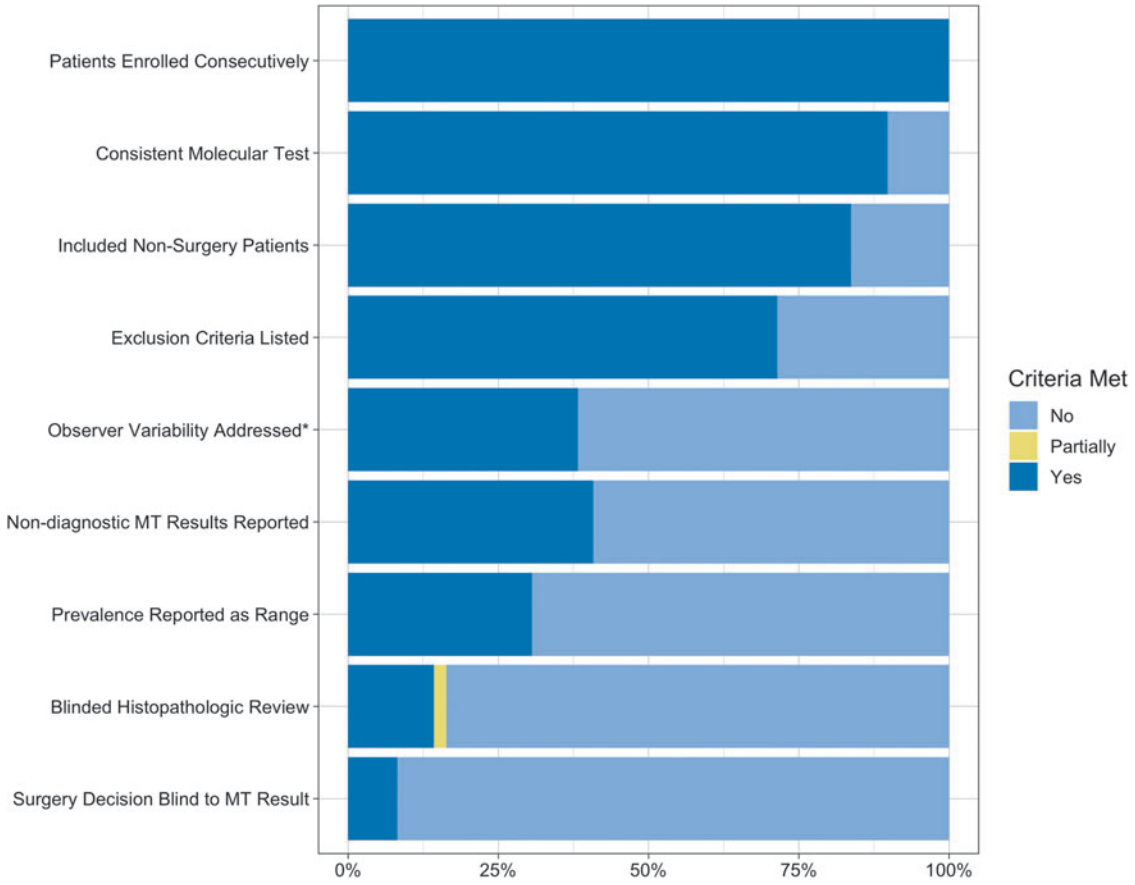

FIG. 2. Potential risk of bias addressed across 49 studies. *Only assessed for studies with multiple institutions or histopathologists reviewing final diagnoses; $n = 34$.

study design (i.e., prospective or retrospective) for Afirma GEC and ThyroSeq v1 and v2 can be found in Supplementary Appendix Tables SA10–SA14. Stratified results were not calculated for Afirma GEC because there was only one study with a prospective design,[71] and they were not calculated for ThyroSeq v3 because there were only two studies with a prospective design.[71,80]

A significant relationship between the inverse of study sample sizes and the reported diagnostic odds ratios would provide evidence of publication bias. There is insufficient evidence of publication bias (Afirma GEC studies, $p = 0.11$; Afirma GSC studies, $p = 0.19$; ThyroSeq v1 and v2 studies, $p = 0.78$; ThyroSeq v3 studies, $p = 0.096$) (Supplementary Appendix Fig. SA2).

Discussion

This is a comprehensive systematic review of clinical validations of molecular tests for thyroid malignancy. We have performed a series of meta-analyses and adapted tools to assess bias within the evaluated studies to contextualize the results of these meta-analyses to further inform future medical and research practices.

We assessed 49 studies of diagnostic accuracy for indeterminate thyroid nodules. Nearly all of the studies followed recommendations to consecutively enroll and separately evaluate the validity of separate molecular tests, and the majority of studies enrolled both patients who went to surgery and patients who did not. These enrollment practices help ensure that the initial patient sample is representative of patients who will undergo a molecular diagnostic test. Most studies reported patient exclusion criteria. A third of studies with multiple histopathologists addressed potential variability between different practitioners' interpretations of the final diagnosis. Less than half of studies reported nondiagnostic test results, which may introduce bias if the true diagnosis of these excluded nodules is differential. Omitting these details hinders evaluations of molecular tests' generalizability across settings.

We offer two major considerations that should be taken in context when interpreting accuracy. First, there was variation in the institutional context (i.e., institutional prevalences and practices such as how many FNAs are performed before ordering a molecular test or which cytological result categories are sent for molecular testing [i.e., indeterminate only, indeterminate + Bethesda V, or all cytological results]). There was also variation in study design, that is, prospective versus retrospective data collection, however, we did not find differences in pooled results when stratifying by study design. This is consistent with a prior systematic review of clinical validations of Afirma GEC, which noted that the initial validation study participants had significantly different nodule characteristics from the patients being tested in practice.[85]

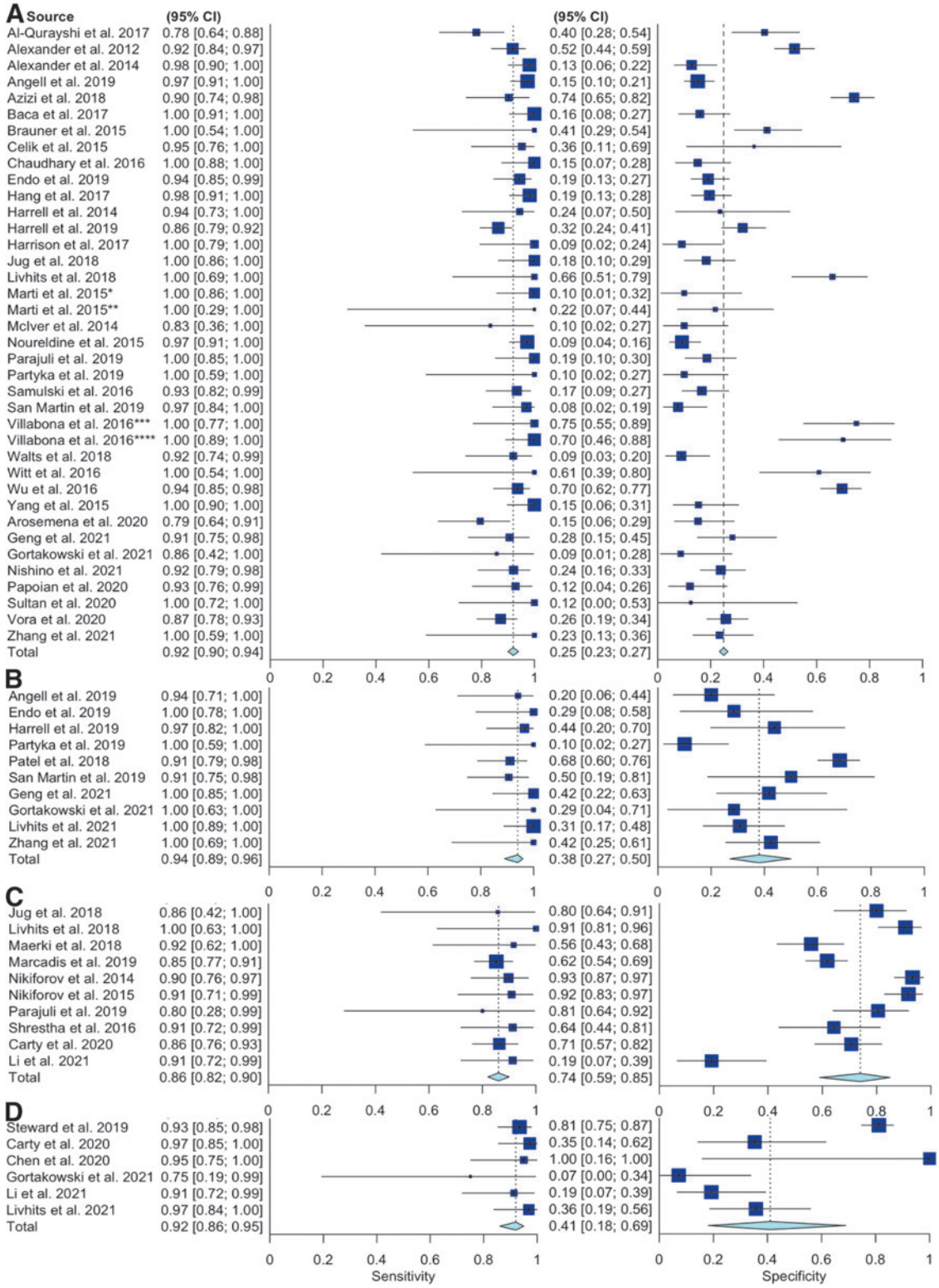

FIG. 3. Sensitivity and specificity of (**A**) Afirma GEC, (**B**) Afirma GSC, (**C**) ThyroSeq v1 and v2, and (**D**) ThyroSeq v3. $N = 49$ studies. For visualization purposes, 0.5 is added to TP, TN, FP, and FN values to avoid division by zero. Lower and upper bounds of confidence interval for summary values are represented by the left and right endpoints of the summary points. *Memorial Sloan Kettering patients; **Mount Sinai Beth Israel patients; ***single FNA performed before molecular test; ****multiple FNAs performed before molecular test. FN, false-negative; FNA, fine needle aspiration; FP, false-positive; GEC, gene expression classifier; TN, true-negative.

This variation enriches the body of literature, but must be accounted for when using these studies to inform clinical management. Given the wide range of institutional prevalences of malignancy and protocols in administering molecular tests (e.g., whether to send a sample for testing after one indeterminate FNA result or two), we found that the sensitivity, specificity, and negative LR measures also vary.

Second, we found that in most studies, the majority of benign molecular test results were never verified by the reference standard (histopathologic diagnosis following surgical resection of the index nodule). Along these same lines, very few studies reported making the decision to send a patient to surgery regardless of molecular test results, which is essential to obtain a representative sample of patients who may undergo a molecular test. This is expected, given the ethical concerns around unnecessary surgery, and indeed, several studies report that the implementation of molecular testing in their clinical practices decreased the incidence of thyroid nodule excision. In addition, the time needed to identify a missed cancer exceeds the study period for many of these studies, and thus, long-term follow-up was not consistently reported. However, the lack of true-negative result verification likely inflates sensitivity.

We found that validations of the four most commonly evaluated molecular tests for indeterminate thyroid nodules—Afirma GEC, Afirma GSC, ThyroSeq v1 and v2, and ThyroSeq v3—did not correct for partial verification bias, either statistically or surgically. Because thyroid malignancies tend to advance slowly,[86,87] it could be months to years before a false-negative result is clinically identified. Thus, the true false-negative rate is unknown. This bias was partially addressed by those studies that only select surgery patients; however, this sample has a higher underlying malignancy prevalence than randomly or consecutively selected samples attributable to other clinical factors.

We also found that the overwhelming majority of studies do not report conducting a blinded histopathologic review, that is, review of a resected specimen without knowledge of the molecular test result. This can expose the diagnostic process to confirmation biases in cases where a diagnostic decision is otherwise on the borderline. These results are consistent with prior systematic reviews of potential biases within studies evaluating Afirma GEC.[21,88] Since the majority of these studies are performed as a retrospective medical records review, it is imperative to use statistical methods to control for underlying biases; however, those methods often require the original patient-level data, which may be personally identifiable. So, these statistical adjustments will be most accurate if they are performed by the original study authors, who do not need to make assumptions about the data.

In addition, we observed a lack of consensus on whether to classify NIFTP findings as benign or malignant; most studies classified NIFTP as malignant, and several evaluated clinical validity in each case. Afirma GEC and ThyroSeq versions 1 and 2 entered the market before NIFTP's reclassification in 2015, so they do not offer any specific diagnostic or management guidance on these nodules. Newer molecular tests, such as Afirma GSC, ThyroSeq v3, or those geared toward microRNA (miRNA) identification similar to ThyraMIR and miRinform, do not advertise the ability to differentiate between NIFTP and benign or malignant nodules. Further study of long-term trajectories of patients with NIFTP will aid in deciding whether to group them more closely with benign or malignant nodules, and future innovations in molecular testing are likely to take this into account.

Our meta-analyses of the four molecular tests with more than three published clinical validations revealed high sensitivities and AUC measures for each test. Our results, in addition to prior systematic reviews of molecular tests for thyroid cancer,[15,88–90] bolster the conclusions of the majority of studies reviewed, including the industry-sponsored studies, which are that molecular tests for indeterminate thyroid nodules have the potential to aid in clinical decision-making; however, solidifying this finding warrants further investigations. For now, given the high level of biases and limitations in the studies evaluated, these results must be interpreted with caution. Future clinical validations of molecular tests must avoid common pitfalls enumerated in this review to evaluate diagnostic molecular tests in a minimally biased manner.

This study has a few limitations. First, different studies have different underlying characteristics (location, patient demographics, observer variability, indeterminate nodule detection rates, study design, patient selection criteria, and cancer prevalence); we find high between-study variation (Supplementary Appendix Tables SA4–SA7). We are unable to assess differences in, for example, patient selection practices from one institution to another, however, we note that these sources of variation may confound the results and introduce additional bias. Notably, while two studies report having received industry sponsorship,[13,70] they report a blinded study design and their results do not appear to differ from the nonsponsored studies.

There are too few studies to perform a secondary subanalysis. Second, there were not enough studies assessing Reveal, ThyraMIR, miRInform, or Quest GMP to conduct meta-analyses. However, these tests are not widely used or available; Afirma GSC and ThyroSeq v3 are the primary diagnostic molecular tests currently utlitized for thyroid cancer, similar to their predecessors, Afirma GEC and ThyroSeq v1 and v2, were before them. Future reviews may benefit from more clinical validations of these tests having been published by the time they are conducted. Third, future reviews may also benefit from head-to-head comparisons of molecular tests, that is, comparisons of different tests' performances on the same cytological sample; only one study in our sample had this design.[58] Fourth, we only included studies that reported diagnostic results, that is, counts of true-negative, true-positive, false-negative, and false-positive results., which may have confounded the results.

Fifth, both because our primary aim was to assess study quality and the reporting of the information needed was limited, we did not perform meta-analyses of test performance stratified by cytological (FNA) results. Finally, we combined ThyroSeq versions 1 and 2 due to version 1 results being combined with version 2 in the studies that did evaluate both, and for studies that noted an institutional switch in which test was used without separating the results, we included their results in the meta-analyses for both tests rather than excluding them from this study. We note, however, that the diagnostic accuracy results of the studies that combined two tests did not differ from the majority of the studies that examined single tests.

Diagnostic molecular testing is a fast-growing market, and oncological tests are projected to claim the largest share of growth over the next decade.[1] Current molecular tests are marketed as rule-out tests and can impact practice by reducing surgeries, but they can also be prohibitively expensive and return indeterminate results. Evaluating independent experiences with all currently and previously commercially available molecular tests for indeterminate thyroid nodules revealed limitations, sources of bias, and gaps in reporting. These biases and gaps in reporting can be addressed in future studies of current and future diagnostic tests.

This systematic review reveals significant sources of bias, which are common among clinical validation studies of commercially available molecular tests for indeterminate thyroid nodules. Meta-analyses of four commonly evaluated and used tests—Afirma GEC, Afirma GSC, ThyroSeq v1 and v2, and ThyroSeq v3—show high sensitivities and AUC measures, which seemingly underscore the suitability and utility of their current use guidelines as a part of thyroid nodule management practices. However, these results must be interpreted in light of high levels of diagnostic review bias and verification bias, in addition to study design limitations. Although molecular tests offer improvements in accuracy over conventional FNA alone, ideally, future validation studies will have prospective designs that err on the side of overinclusion to ensure accurate perceptions of the value added by molecular testing. Moreover, the role of patient decision-making and decision-making aids in the use of these tools should be considered and evaluated.

Authors' Contributions

C.D.: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft, writing—review and editing, and visualization; V.V.: conceptualization, methodology, data curation, validation, and writing—review and editing; M.S.J.: writing—review and editing and methodology; A.T.: writing—review and editing; T.W.: writing—review and editing; S.G.: writing—review and editing; N.M.: writing—review and editing, methodology, and software; C.C.L.: conceptualization, methodology, supervision, project administration, writing—review and editing, and funding acquisition.

Acknowledgments

We thank Lisa Philpott and Melissa Lydston at the Massachusetts General Hospital Treadwell Library for constructing the initial search query and performing the academic database search and for providing key feedback on our initial search strategy.

Author Disclosure Statement

The authors have nothing to disclose.

Funding Information

This work was supported by NIH/NCI R37 CA231957 (C.C.L.). The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the article; or decision to submit the article for publication.

Supplementary Material

Supplementary Appendix Methods
Supplementary Appendix Figure SA1
Supplementary Appendix Figure SA2
Supplementary Appendix Table SA1
Supplementary Appendix Table SA2
Supplementary Appendix Table SA3
Supplementary Appendix Table SA4
Supplementary Appendix Table SA5
Supplementary Appendix Table SA6
Supplementary Appendix Table SA7
Supplementary Appendix Table SA8
Supplementary Appendix Table SA9
Supplementary Appendix Table SA10
Supplementary Appendix Table SA11
Supplementary Appendix Table SA12
Supplementary Appendix Table SA13
Supplementary Appendix Table SA14

References

1. Grand View Research. North American Molecular Diagnostics Market Size, Share & Trends Analysis Report by Technology, by Application (Oncology, CVD), by Test Location (PoC, OTC), by Product (Instruments, Reagents), and Segment Forecasts, 2018–2025. 2018. Available at: https://www.grandviewresearch.com/industry-analysis/north-america-molecular-diagnostics-market
2. Mukherjee S, Fountain G, Stalker M, et al. The 'straight to test' initiative reduces both diagnostic and treatment waiting times for colorectal cancer: Outcomes after 2years. Colorectal Dis 2010;12(10):e250–e254; doi: 10.1111/j.1463-1318.2009.02182.x
3. Renwick L, Hardie A, Girvan EK, et al. Detection of meticillin-resistant Staphylococcus aureus and Panton-Valentine leukocidin directly from clinical samples and the development of a multiplex assay using real-time polymerase chain reaction. Eur J Clin Microbiol Infect Dis 2008;27(9):791–796; doi: 10.1007/s10096-008-0503-9
4. Fang C, Otero HJ, Greenberg D, et al. Cost-utility analyses of diagnostic laboratory tests: A systematic review. Value Health 2011;14(8):1010–1018; doi: 10.1016/j.jval.2011.05.044
5. Nikiforov YE, Steward DL, Robinson-Smith TM, et al. Molecular testing for mutations in improving the fine-needle aspiration diagnosis of thyroid nodules. J Clin Endocrinol Metab 2009;94(6):2092–2098; doi: 10.1210/jc.2009–0247
6. Qaseem A, Alguire P, Dallas P, et al. Appropriate use of screening and diagnostic tests to foster high-value, cost-conscious care. Ann Intern Med 2012;156(2):147–149; doi: 10.7326/0003-4819-156-2-201201170-00011
7. Hunink MM, Weinstein MC, Wittenberg E, et al. Decision Making in Health and Medicine: Integrating Evidence and Values. Cambridge: Cambridge University Press; 2014.
8. Swets JA, Dawes RM, Monahan J. Psychological science can improve diagnostic decisions. Psychol Sci Public Interest 2000;1(1):1–26; doi: 10.1111/1529-1006.001
9. Deverka P, Messner DA, McCormack R, et al. Generating and evaluating evidence of the clinical utility of molecular diagnostic tests in oncology. Genet Med 2016;18(8):780–787; doi: 10.1038/gim.2015.162.
10. Nikiforova MN, Nikiforov YE. Molecular diagnostics and predictors in thyroid cancer. Thyroid 2009;19(12):1351–1361; doi: 10.1089/thy.2009.0240

11. Crippa S, Mazzucchelli L. The Bethesda system for reporting thyroid fine-needle aspiration specimens. Am J Clin Pathol 2010;134(2):343–345; doi: 10.1309/ajcpxm9wirq8jzbj
12. Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: The American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. Thyroid 2016;26(1):1–133; doi: 10.1089/thy.2015.0020
13. Alexander EK, Kennedy GC, Baloch ZW, et al. Preoperative diagnosis of benign thyroid nodules with indeterminate cytology. N Engl J Med 2012;367(8):705–715; doi: 10.1056/NEJMoa1203208
14. Nicholson KJ, Yip L. An update on the status of molecular testing for the indeterminate thyroid nodule and risk stratification of differentiated thyroid cancer. Cur Opin Oncol 2018;30(1):8–15; doi: 10.1097/CCO.0000000000000414
15. Vargas-Salas S, Martinez JR, Urra S, et al. Genetic testing for indeterminate thyroid cytology: Review and meta-analysis. Endocr Relat Cancer 2018;25(3):R163–R177; doi: 10.1530/ERC-17-0405
16. Mitchell J, Yip L. Decision making in indeterminate thyroid nodules and the role of molecular testing. Surg Clin North Am 2019;99(4):587–598; doi: 10.1016/j.suc.2019.04.002
17. Vuong HG, Nguyen TPX, Hassell LA, et al. Diagnostic performances of the Afirma gene sequencing classifier in comparison with the gene expression classifier: A meta-analysis. Cancer Cytopathol 2021;129(3):182–189; doi: 10.1002/cncy.22332
18. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ 2021;372(8286):n71; doi: 10.1136/bmj.n71.
19. Mallett S, Halligan S, Thompson M, et al. Interpreting diagnostic accuracy studies for patient care. BMJ 2012;345(7871):e3999; doi: 10.1136/bmj.e3999
20. Umemneku Chikere CM, Wilson K, Graziadio S, et al. Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard—An update. PLoS One 2019;14(10):e0223832; doi: 10.1371/journal.pone.0223832
21. Duh Q-Y, Busaidy NL, Rahilly-Tierney C, et al. A systematic review of the methods of diagnostic accuracy studies of the Afirma gene expression classifier. Thyroid 2017;27(10):1215–1222; doi: 10.1089/thy.2016.0656
22. Kosinski AS, Barnhart HX. A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present. Stat Med 2003;22(17):2711–2721; doi: 10.1002/sim.1517
23. Collins J, Huynh M. Estimation of diagnostic test accuracy without full verification: A review of latent class methods. Stat Med 2014;33(24):4141–4169; doi: 10.1002/sim.6218
24. Zhou X-H. Correcting for verification bias in studies of a diagnostic test's accuracy. Stat Methods Med Res 1998;7(4):337–353; doi: 10.1177/096228029800700403
25. Alonzo TA. Verification bias-impact and methods for correction when assessing accuracy of diagnostic tests. Revstat Stat J 2014;12(1):67–83; doi: 10.57805/revstat.v12i1.144
26. Reitsma JB, Rutjes AW, Khan KS, et al. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol 2009;62(8):797–806; doi: 10.1016/j.jclinepi.2009.02.005
27. Whiting P, Rutjes AW, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy: A systematic review. Ann Intern Med 2004;140(3):189–202; doi: 10.7326/0003-4819-140-3-200402030-00010
28. R Core Team. R: A Language and Environment for Statistical Computing. . R Foundation for Statistical Computing: Vienna; 2013.
29. Glas AS, Lijmer JG, Prins MH, et al. The diagnostic odds ratio: A single indicator of test performance. J Clin Epidemiol 2003;56(11):1129–1135; doi: 10.1016/s0895-4356(03)00177-x
30. Cochran WG. The comparison of percentages in matched samples. Biometrika 1950;37(3/4):256–266; doi: https://psycnet.apa.org/doi/10.1093/biomet/37.3-4.256
31. Schaarschmidt F. bdpv: Inference and Design for Predictive Values in Diagnostic Tests. 2019. Available at: https://cran.r-project.org/package=bdpv
32. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26(4):404–413; doi: 10.1093/biomet/26.4.404
33. Reitsma JB, Glas AS, Rutjes AW, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58(10):982–990; doi: 10.1016/j.jclinepi.2005.02.022
34. Davison AC, Hinkley DV. Bootstrap Methods and Their Application. Cambridge: Cambridge University Press; 1997.
35. Noma H, Matsushima Y. Confidence interval for the AUC of SROC curve and some related methods using bootstrap for meta-analysis of diagnostic accuracy studies. Commun Stat Case Stud Data Anal Appl 2021;7(3):344–358; doi: 10.13039/501100001691
36. Al-Qurayshi Z, Deniwar A, Thethi T, et al. Association of malignancy prevalence with test properties and performance of the gene expression classifier in indeterminate thyroid nodules. JAMA Otolaryngol Head Neck Surg 2017;143(4):403–408; doi: 10.1001/jamaoto.2016.3526
37. Alexander EK, Schorr M, Klopper J, et al. Multicenter clinical experience with the Afirma gene expression classifier. J Clin Endocrinol Metab 2014;99(1):119–125; doi: 10.1210/jc.2013–2482
38. Angell TE, Heller HT, Cibas ES, et al. Independent comparison of the Afirma genomic sequencing classifier and gene expression classifier for cytologically indeterminate thyroid nodules. Thyroid 2019;29(5):650–656; doi: 10.1089/thy.2018.0726
39. Azizi G, Keller JM, Mayo ML, et al. Shear wave elastography and Afirma gene expression classifier in thyroid nodules with indeterminate cytology: A comparison study. Endocrine 2018;59(3):573–584; doi: 10.1007/s12020-017-1509-9
40. Baca SC, Wong KS, Strickland KC, et al. Qualifiers of atypia in the cytologic diagnosis of thyroid nodules are associated with different Afirma gene expression classifier results and clinical outcomes. Cancer Cytopathol 2017;125(5):313–322; doi: 10.1002/cncy.21827
41. Brauner E, Holmes BJ, Krane JF, et al. Performance of the Afirma gene expression classifier in hurthle cell thyroid nodules differs from other indeterminate thyroid nodules. Thyroid 2015;25(7):789–796; doi: 10.1089/thy.2015.0049
42. Celik B, Whetsell CR, Nassar A. Afirma GEC and thyroid lesions: An institutional experience. Diagn Cytopathol 2015;43(12):966–970; doi: 10.1002/dc.23378
43. Chaudhary S, Hou Y, Shen R, et al. Impact of the Afirma gene expression classifier result on the surgical manage-

ment of thyroid nodules with category III/IV cytology and its correlation with surgical outcome. Acta Cytol 2016; 60(3):205–210; doi: 10.1159/000446797
44. Endo M, Nabhan F, Porter K, et al. Afirma gene sequencing classifier compared with gene expression classifier in indeterminate thyroid nodules. Thyroid 2019;29(8):1115–1124; doi: 10.1089/thy.2018.0733
45. Hang JF, Westra WH, Cooper DS, et al. The impact of noninvasive follicular thyroid neoplasm with papillary-like nuclear features on the performance of the Afirma gene expression classifier. Cancer Cytopathol 2017;125(9):683–691; doi: 10.1002/cncy.21879
46. 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(4):364–369; doi: 10.4158/EP13330.OR
47. Harrell RM, Eyerly-Webb SA, Golding AC, et al. Statistical comparison of Afirma GSC and Afirma GEC outcomes in a community endocrine surgical practice: Early findings. Endocr Pract 2019;25(2):161–164; doi: 10.4158/EP-2018-0395
48. Harrison G, Sosa JA, Jiang X. Evaluation of the Afirma gene expression classifier in repeat indeterminate thyroid nodules. Arch Pathol Lab Med 2017;141(7):985–989; doi: 10.5858/arpa.2016-0328-OA
49. Jug RC, Datto MB, Jiang XS. Molecular testing for indeterminate thyroid nodules: Performance of the Afirma gene expression classifier and ThyroSeq panel. Cancer Cytopathol 2018;126(7):471–480; doi: 10.1002/cncy.21993
50. Livhits MJ, Kuo EJ, Leung AM, et al. Gene expression classifier vs targeted next-generation sequencing in the management of indeterminate thyroid nodules. J Clin Endocrinol Metab 2018;103(6):2261–2268; doi: 10.1210/jc.2017–02754
51. Marti JL, Avadhani V, Donatelli LA, et al. Wide inter-institutional variation in performance of a molecular classifier for indeterminate thyroid nodules. Ann Surg Oncol 2015;22(12):3996–4001; doi: 10.1245/s10434-015-4486-3
52. 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(11):4069–4077; doi: 10.1210/jc.2013–3584
53. Noureldine SI, Olson MT, Agrawal N, et al. Effect of gene expression classifier molecular testing on the surgical decision-making process for patients with thyroid nodules. JAMA Otolaryngol Head Neck Surg 2015;141(12):1082–1088; doi: 10.1001/jamaoto.2015.2708
54. Parajuli S, Jug R, Ahmadi S, et al. Hurthle cell predominance impacts results of Afirma gene expression classifier and ThyroSeq molecular panel performance in indeterminate thyroid nodules. Diagn Cytopathol 2019;47(11):1177–1183; doi: 10.1002/dc.24290
55. Samulski TD, LiVolsi VA, Wong LQ, et al. Usage trends and performance characteristics of a ''gene expression classifier'' in the management of thyroid nodules: An institutional experience. Diagn Cytopathol 2016;44(11):867–873; doi: 10.1002/dc.23559
56. San Martin VT, Lawrence L, Bena J, et al. Real-world comparison of Afirma GEC and GSC for the assessment of cytologically indeterminate thyroid nodules. J Clin Endocrinol Metab 2020;105(3):dgz099; doi: 10.1210/clinem/dgz099
57. Villabona CV, Mohan V, Arce KM, et al. Utility of ultrasound versus gene expression classifier in thyroid nodules with atypia of undetermined significance. Endocr Pract 2016;22(10):1199–1203; doi: 10.4158/EP161231.OR
58. Walts AE, Sacks WL, Wu HH, et al. A retrospective analysis of the performance of the RosettaGX((R)) Reveal thyroid miRNA and the Afirma Gene Expression Classifiers in a cohort of cytologically indeterminate thyroid nodules. Diagn Cytopathol 2018;46(11):901–907; doi: 10.1002/dc.23980
59. Witt RL. Outcome of thyroid gene expression classifier testing in clinical practice. Laryngoscope 2016;126(2):524–527; doi: 10.1002/lary.25607
60. Wu JX, Young S, Hung ML, et al. Clinical factors influencing the performance of gene expression classifier testing in indeterminate thyroid nodules. Thyroid 2016;26(7):916–922; doi: 10.1089/thy.2015.0505
61. Yang SE, Sullivan PS, Zhang J, et al. Has Afirma gene expression classifier testing refined the indeterminate thyroid category in cytology? Cancer Cytopathol 2016;124(2):100–109; doi: 10.1002/cncy.21624
62. Arosemena M, Thekkumkattil A, Valderrama ML, et al. American Thyroid Association sonographic risk and Afirma gene expression classifier alone and in combination for the diagnosis of thyroid nodules with Bethesda category III cytology. Thyroid 2020;30(11):1613–1619; doi: 10.1089/thy.2019.0673
63. Geng Y, Aguilar-Jakthong JS, Moatamed NA. Comparison of Afirma Gene Expression Classifier with Gene Sequencing Classifier in indeterminate thyroid nodules: A single-institutional experience. Cytopathology 2021;32(2):187–191; doi: 10.1111/cyt.12920
64. Gortakowski M, Feghali K, Osakwe I. Single institution experience with Afirma and Thyroseq testing in indeterminate thyroid nodules. Thyroid 2021;31(9):1376–1382; doi: 10.1089/thy.2020.0801
65. Nishino M, Mateo R, Kilim H, et al. Repeat fine needle aspiration cytology refines the selection of thyroid nodules for Afirma gene expression classifier testing. Thyroid 2021;31(8):1253–1263; doi: 10.1089/thy.2020.0969
66. Papoian V, Rosen JE, Lee W, et al. Differentiated thyroid cancer and Hashimoto thyroiditis: Utility of the Afirma gene expression classifier. J Surg Oncol 2020;121(7):1053–1057; doi: 10.1002/jso.25875
67. Sultan R, Levy S, Sulanc E, et al. Utility of Afirma gene expression classifier for evaluation of indeterminate thyroid nodules and correlation with ultrasound risk assessment: Single institutional experience. Endocr Pract 2020;26(5):543–551; doi: 10.4158/EP-2019-0350
68. Vora A, Holt S, Haque W, et al. Long-term outcomes of thyroid nodule AFIRMA GEC testing and literature review: An institutional experience. Otolaryngol Head Neck Surg 2020;162(5):634–640; doi: 10.1177/0194599820911718
69. Zhang L, Smola B, Lew M, et al. Performance of Afirma genomic sequencing classifier vs gene expression classifier in Bethesda category III thyroid nodules: An institutional experience. Diagn Cytopathol 2021;49(8):921–927; doi: 10.1002/dc.24765
70. Patel KN, Angell TE, Babiarz J, et al. Performance of a genomic sequencing classifier for the preoperative diagnosis of cytologically indeterminate thyroid nodules. JAMA Surg 2018;153(9):817–824; doi: 10.1001/jamasurg.2018.1153

71. Livhits MJ, Zhu CY, Kuo EJ, et al. Effectiveness of molecular testing techniques for diagnosis of indeterminate thyroid nodules: A randomized clinical trial. JAMA Oncol 2021;7(1):70–77; doi: 10.1001/jamaoncol.2020.5935
72. Partyka KL, Trevino K, Randolph ML, et al. Risk of malignancy and neoplasia predicted by three molecular testing platforms in indeterminate thyroid nodules on fine-needle aspiration. Diagn Cytopathol 2019;47(9):853–862; doi: 10.1002/dc.24250
73. Maerki J, Klein M, Chau K, et al. Determining the molecular test for indeterminate thyroid nodules best suited for our practice: A quality assurance study. Diagn Cytopathol 2019;47(4):259–267; doi: 10.1002/dc.24091
74. Marcadis AR, Valderrabano P, Ho AS, et al. Interinstitutional variation in predictive value of the ThyroSeq v2 genomic classifier for cytologically indeterminate thyroid nodules. Surgery 2019;165(1):17–24; doi: 10.1016/j.surg.2018.04.062
75. Nikiforov YE, Carty SE, Chiosea SI, et al. Impact of the multi-gene ThyroSeq next-generation sequencing assay on cancer diagnosis in thyroid nodules with atypia of undetermined significance/follicular lesion of undetermined significance cytology. Thyroid 2015;25(11):1217–1223; doi: 10.1089/thy.2015.0305
76. 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(23):3627–3634; doi: 10.1002/cncr.29038
77. Shrestha RT, Evasovich MR, Amin K, et al. Correlation between histological diagnosis and mutational panel testing of thyroid nodules: A two-year institutional experience. Thyroid 2016;26(8):1068–1076; doi: 10.1089/thy.2016.0048
78. Carty SE, Ohori NP, Hilko DA, et al. The clinical utility of molecular testing in the management of thyroid follicular neoplasms (Bethesda IV nodules). Ann Surg 2020;272(4):621–627; doi: 10.1097/SLA.0000000000004130
79. Jug R, Parajuli S, Ahmadi S, et al. Negative results on thyroid molecular testing decrease rates of surgery for indeterminate thyroid nodules. Endocr Pathol 2019;30(2):134–137; doi: 10.1007/s12022-019-9571-x
80. Steward DL, Carty SE, Sippel RS, et al. Performance of a multigene genomic classifier in thyroid nodules with indeterminate cytology: A prospective blinded multicenter study. JAMA Oncol 2019;5(2):204–212; doi: 10.1001/jamaoncol.2018.4616
81. Chen T, Gilfix BM, Rivera J, et al. The role of the ThyroSeq v3 molecular test in the surgical management of thyroid nodules in the Canadian public health care setting. Thyroid 2020;30(9):1280–1287; doi: 10.1089/thy.2019.0539
82. Li W, Justice-Clark T, Cohen MB. The utility of ThyroSeq® in the management of indeterminate thyroid nodules by fine-needle aspiration. Cytopathology 2021;32(4):505–512; doi: 10.1111/cyt.12981
83. Labourier E, Shifrin A, Busseniers AE, et al. Molecular testing for miRNA, mRNA, and DNA on fine-needle aspiration improves the preoperative diagnosis of thyroid nodules with indeterminate cytology. J Clin Endocrinol Metab 2015;100(7):2743–2750; doi: 10.1210/jc.2015-1158
84. Valderrabano P, Leon ME, Centeno BA, et al. Institutional prevalence of malignancy of indeterminate thyroid cytology is necessary but insufficient to accurately interpret molecular marker tests. Eur J Endocrinol 2016;174(5):621–629; doi: 10.1530/EJE-15-1163
85. Valderrabano P, Hallanger-Johnson JE, Thapa R, et al. Comparison of postmarketing findings vs the initial clinical validation findings of a thyroid nodule gene expression classifier: A systematic review and meta-analysis. JAMA Otolaryngol Head Neck Surg 2019;145(9):783–792; doi: 10.1001/jamaoto.2019.1449
86. Ito Y, Miyauchi A, Kihara M, et al. Patient age is significantly related to the progression of papillary microcarcinoma of the thyroid under observation. Thyroid 2014;24(1):27–34; doi: 10.1089/thy.2013.0367
87. Sugitani I, Toda K, Yamada K, et al. Three distinctly different kinds of papillary thyroid microcarcinoma should be recognized: Our treatment strategies and outcomes. World J Surg 2010;34(6):1222–1231; doi: 10.1007/s00268-009-0359-x
88. Liu Y, Pan B, Xu L, et al. The diagnostic performance of Afirma gene expression classifier for the indeterminate thyroid nodules: A meta-analysis. Biomed Res Int 2019;2019:7150527; doi: 10.1155/2019/7150527
89. Borowczyk M, Szczepanek-Parulska E, Olejarz M, et al. Evaluation of 167 Gene Expression Classifier (GEC) and ThyroSeq v2 diagnostic accuracy in the preoperative assessment of indeterminate thyroid nodules: Bivariate/HROC meta-analysis. Endocr Pathol 2019;30(1):8–15; doi: 10.1007/s12022-018-9560-5
90. Santhanam P, Khthir R, Gress T, et al. Gene expression classifier for the diagnosis of indeterminate thyroid nodules: A meta-analysis. Med Oncol 2016;33(2):14; doi: 10.1007/s12032-015-0727-3
91. Campbell J, Klugar M, Ding S, et al. The Systematic Review of Studies of Diagnostic Test Accuracy. Joanna Briggs Institute Reviewers' Manual. The Joanna Briggs Institute: Adelaide, South Australia; 2015; pp. 1–46.
92. Santaguida PL, Riley CM, Matchar DB. Assessing risk of bias as a domain of quality in medical test studies. J Gen Intern Med 2012;27(Suppl 1):33–38; doi: 10.1007/s11606-012-2030-8

Address correspondence to:
Carrie Cunningham Lubitz, MD, MPH
Harvard Medical School
Department of Surgery
Massachusetts General Hospital
Section of Endocrine Surgery
55 Fruit St., Yawkey 7B
Boston, MA 02114
USA

E-mail: clubitz@mgh.harvard.edu