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 Mergaio,1,3 and Carrie Cunningham Lubitz1,2

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
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. While diagnostic tests have shown benefits by reducing unnecessary surgeries and hospitalizations, improving the accuracy of decisions made by practitioners, 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.

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. 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 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 result (i.e., a missed cancer), molecular tests (MTs) for thyroid cancer were developed with a high sensitivity to rule out a false-negative result. For instance, a false positive may result in unnecessary treatment; a false negative may result in a missed cancer.

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. While Bethesda III and IV, in the six-tiered Bethesda System for Reporting Thyroid Cytology, are considered indeterminate FNA results. The American Thyroid Association recommended the use of molecular tests for indeterminate thyroid nodules in clinical practice in 2015. 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. Veracyte developed the first widely used molecular test, the Afirma GEC molecular test in 2012. The test was widely adopted after successful clinical trials. Other companies soon followed suit.

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 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. 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,” “cytologic 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,” “non-medullary 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, and the process of performing one such study is characterized elsewhere. 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 and statistically corrected. 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. 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 test29 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, evaluated the clinical validity of Afirma GEC; nine, evaluated Afirma GSC. One study 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 evaluated ThyroSeq v1 or v2 (version 1 and version 2 results were combined in the studies that did evaluate both), and five studies evaluated ThyroSeq v3 (a significant expansion of previous versions). One study evaluated both ThyroSeq v2 and 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 studies evaluated Rosetta GX Reveal, two evaluated miRNeform, one evaluated TyhraMIR, and one evaluated Quest Diagnostics’ gene mutation panel (GMP).

Of the 49 studies included, 14 compared 2 diagnostic tests and 26 compared 3 diagnostic tests. Thirty-nine studies used a retrospective design, nine studies used a prospective design, and one study combined retrospective and prospective cohorts. Twenty-five studies performed a single FNA before molecular testing and two 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>
<thead>
<tr>
<th>Biases</th>
<th>Description</th>
<th>Impact on accuracy</th>
<th>Preventative measures</th>
<th>Impact on thyroid cancer diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td>Eligible participants are not enrolled randomly or consecutively</td>
<td>Depends on who is not represented</td>
<td>Ensure that participants are enrolled randomly or consecutively, report participant exclusion criteria</td>
<td>If the study population does not represent the underlying characteristics of the available population, the results may not be generalizable to another population. 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. When multiple pathologists interpret histopathologic results, there may be differences in their conclusions; this can bias the assessed accuracy of the molecular test. Molecular tests for thyroid cancer return nondiagnostic results frequently; to determine the clinical utility of a molecular test, nondiagnostic results must be reported. Histopathologic diagnoses of thyroid nodules are interpreted after molecular tests in prospective studies and retrospective studies where the test was conducted during the clinical timeline 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. 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. 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.</td>
</tr>
</tbody>
</table>

| 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 |                                                                                                                                                                  |

| 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 |                                                                                                                                                                  |

| 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                                                                 |                                                                                                                                                                  |

| Limitations                   | 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                                      |                                                                                                                                                                  |

| 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                           |                                                                                                                                                                  |

| 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               |                                                                                                                                                                  |

---

\(^a\)Histopathologic diagnosis (generally, the “gold standard” test revealing ground truth).

\(^b\)Molecular diagnostic test (generally, the test being evaluated).

Adapted from Campbell et al.\(^1\) and Santaguida et al.\(^2\)
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 \((n = 10; p = 0.54)\), 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>
<thead>
<tr>
<th>Authors (year)</th>
<th>Ref.</th>
<th>Country</th>
<th>Sites</th>
<th>Study design</th>
<th>Molecular test</th>
<th>Patients</th>
<th>Nodules</th>
<th>FNAs before MT</th>
<th>NIFTP designation</th>
<th>Industry sponsorship</th>
<th>Quality criteria met (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al-Qurayshi et al (2017)</td>
<td>36</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>145</td>
<td>154</td>
<td>1</td>
<td>—</td>
<td>No</td>
<td>50</td>
</tr>
<tr>
<td>Alexander et al (2012)</td>
<td>43</td>
<td>United States</td>
<td>49</td>
<td>Prospective</td>
<td>AGEC</td>
<td>249</td>
<td>265</td>
<td>1</td>
<td>Pre-2016</td>
<td>Yes</td>
<td>78</td>
</tr>
<tr>
<td>Alexander et al (2014)</td>
<td>57</td>
<td>United States</td>
<td>5</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>339</td>
<td>339</td>
<td>1</td>
<td>Pre-2016</td>
<td>No</td>
<td>56</td>
</tr>
<tr>
<td>Angell et al (2019)</td>
<td>34</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>563</td>
<td>486</td>
<td>&gt;1</td>
<td>—</td>
<td>Malignant</td>
<td>No</td>
</tr>
<tr>
<td>Arosemena et al (2020)</td>
<td>62</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>117</td>
<td>126</td>
<td>Unreported</td>
<td>Malignant</td>
<td>No</td>
<td>44</td>
</tr>
<tr>
<td>Azizi et al (2018)</td>
<td>59</td>
<td>United States</td>
<td>1</td>
<td>Prospective</td>
<td>AGEC</td>
<td>151</td>
<td>151</td>
<td>&gt;1</td>
<td>Benign</td>
<td>No</td>
<td>44</td>
</tr>
<tr>
<td>Baca et al (2017)</td>
<td>30</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>219</td>
<td>227</td>
<td>&gt;1</td>
<td>Malignant</td>
<td>No</td>
<td>50</td>
</tr>
<tr>
<td>Brauner et al (2015)</td>
<td>41</td>
<td>United States</td>
<td>3</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>72</td>
<td>72</td>
<td>1</td>
<td>Pre-2016</td>
<td>No</td>
<td>67</td>
</tr>
<tr>
<td>Carty et al (2020)</td>
<td>78</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>TSv2 and v3</td>
<td>240</td>
<td>1</td>
<td>Both</td>
<td>No</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>Celik et al (2015)</td>
<td>42</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>66</td>
<td>66</td>
<td>1</td>
<td>Pre-2016</td>
<td>No</td>
<td>50</td>
</tr>
<tr>
<td>Chaudhary et al (2016)</td>
<td>43</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>158</td>
<td>158</td>
<td>&gt;1</td>
<td>Unclear</td>
<td>No</td>
<td>56</td>
</tr>
<tr>
<td>Chen et al (2020)</td>
<td>81</td>
<td>Canada</td>
<td>1</td>
<td>Retrospective</td>
<td>TSv3</td>
<td>50</td>
<td>50</td>
<td>1</td>
<td>Malignant</td>
<td>No</td>
<td>44</td>
</tr>
<tr>
<td>Celik et al (2015)</td>
<td>42</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>66</td>
<td>66</td>
<td>1</td>
<td>Pre-2016</td>
<td>No</td>
<td>50</td>
</tr>
<tr>
<td>Harrell et al (2014)</td>
<td>46</td>
<td>United States</td>
<td>1</td>
<td>Prospective</td>
<td>AGEC</td>
<td>58</td>
<td>58</td>
<td>1</td>
<td>Pre-2016</td>
<td>No</td>
<td>56</td>
</tr>
<tr>
<td>Harrell et al (2019)</td>
<td>47</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>481</td>
<td>1</td>
<td>Malignant</td>
<td>No</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>Harrison et al (2017)</td>
<td>48</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>110</td>
<td>115</td>
<td>&gt;1</td>
<td>—</td>
<td>No</td>
<td>44</td>
</tr>
<tr>
<td>Jug et al (2018)</td>
<td>49</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>TSv1 and v2</td>
<td>91</td>
<td>97</td>
<td>1</td>
<td>Both</td>
<td>No</td>
<td>56</td>
</tr>
<tr>
<td>Li et al (2021)</td>
<td>82</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>TSv2 and v3</td>
<td>202</td>
<td>1</td>
<td>Benign</td>
<td>No</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Livhits et al (2018)</td>
<td>50</td>
<td>United States</td>
<td>1</td>
<td>Prospective</td>
<td>AGEC</td>
<td>70</td>
<td>70</td>
<td>1</td>
<td>Malignant</td>
<td>No</td>
<td>78</td>
</tr>
<tr>
<td>Livhits et al (2021)</td>
<td>71</td>
<td>United States</td>
<td>1</td>
<td>Prospective</td>
<td>AGSC</td>
<td>189</td>
<td>201</td>
<td>1</td>
<td>Benign</td>
<td>No</td>
<td>67</td>
</tr>
<tr>
<td>Maerkki et al (2019)</td>
<td>73</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>Quest GMP</td>
<td>157</td>
<td>171</td>
<td>1</td>
<td>—</td>
<td>Benign</td>
<td>No</td>
</tr>
<tr>
<td>Marcadis et al (2019)</td>
<td>74</td>
<td>United States</td>
<td>4</td>
<td>Retrospective</td>
<td>TSv2 and v3</td>
<td>266</td>
<td>273</td>
<td>—</td>
<td>Both</td>
<td>No</td>
<td>56</td>
</tr>
<tr>
<td>Marti et al (2015)</td>
<td>51,c</td>
<td>United States</td>
<td>2</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>94</td>
<td>94</td>
<td>—</td>
<td>Pre-2016</td>
<td>No</td>
<td>44</td>
</tr>
<tr>
<td>McIver et al (2014)</td>
<td>72</td>
<td>United States</td>
<td>1</td>
<td>Prospective</td>
<td>AGEC</td>
<td>72</td>
<td>72</td>
<td>1</td>
<td>Pre-2016</td>
<td>No</td>
<td>75</td>
</tr>
<tr>
<td>Nikiforov et al (2014)</td>
<td>76</td>
<td>United States</td>
<td>1</td>
<td>Both</td>
<td>TSv2</td>
<td>143</td>
<td>143</td>
<td>—</td>
<td>Pre-2016</td>
<td>No</td>
<td>50</td>
</tr>
<tr>
<td>Nikiforov et al (2015)</td>
<td>75</td>
<td>United States</td>
<td>1</td>
<td>Prospective</td>
<td>TSv2</td>
<td>441</td>
<td>465</td>
<td>1</td>
<td>Pre-2016</td>
<td>No</td>
<td>50</td>
</tr>
<tr>
<td>Nishino et al (2021)</td>
<td>65</td>
<td>United States</td>
<td>1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>360</td>
<td>370</td>
<td>&gt;1</td>
<td>Malignant</td>
<td>No</td>
<td>67</td>
</tr>
</tbody>
</table>

(continued)
Table 2. (Continued)

<table>
<thead>
<tr>
<th>Authors (year) Ref.</th>
<th>Country</th>
<th>Sites</th>
<th>Study design</th>
<th>Molecular test</th>
<th>Patients</th>
<th>Nodules before MT</th>
<th>NIFTP designation</th>
<th>Industry sponsorship</th>
<th>Quality criteria met (%) a,b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noureldine et al (2015) 53</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>273</td>
<td>274</td>
<td>&gt;1</td>
<td>Pre-2016</td>
<td>No</td>
<td>50</td>
</tr>
<tr>
<td>Papoian et al (2020) 66</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>69</td>
<td>69</td>
<td>1</td>
<td>Unclear</td>
<td>No</td>
<td>44</td>
</tr>
<tr>
<td>Parajuli et al (2019) 54</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>199</td>
<td>202</td>
<td>&gt;1</td>
<td>Benign</td>
<td>No</td>
<td>50</td>
</tr>
<tr>
<td>Partyka et al (2019) 72</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>TSv1 and v2</td>
<td>81</td>
<td>81</td>
<td>—</td>
<td>Benign</td>
<td>No</td>
<td>44</td>
</tr>
<tr>
<td>Patel et al (2018) 70</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>183</td>
<td>190</td>
<td>1</td>
<td>Malignant</td>
<td>Yes</td>
<td>67</td>
</tr>
<tr>
<td>Samulski et al (2016) 52</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>294</td>
<td>294</td>
<td>&gt;1</td>
<td>Malignant</td>
<td>No</td>
<td>63</td>
</tr>
<tr>
<td>San Martin et al (2020) 56</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>174</td>
<td>178</td>
<td>1</td>
<td>—</td>
<td>No</td>
<td>50</td>
</tr>
<tr>
<td>Shrestha et al (2016) 77</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>TSv1 and v2</td>
<td>261</td>
<td>261</td>
<td>1</td>
<td>—</td>
<td>No</td>
<td>38</td>
</tr>
<tr>
<td>Steward et al (2019) 80</td>
<td>United States, Singapore 10</td>
<td>Prospective</td>
<td>TSv3</td>
<td>257</td>
<td>286</td>
<td>1</td>
<td>Malignant</td>
<td>No</td>
<td>100</td>
</tr>
<tr>
<td>Sultan et al (2020) 67</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>98</td>
<td>101</td>
<td>&gt;1</td>
<td>Both</td>
<td>No</td>
<td>33</td>
</tr>
<tr>
<td>Valderrabano et al (2016) 84</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>miRInform</td>
<td>105</td>
<td>116</td>
<td>&gt;1</td>
<td>—</td>
<td>No</td>
<td>75</td>
</tr>
<tr>
<td>Villabona et al (2016) 77</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>48</td>
<td>48</td>
<td>1 and &gt;1</td>
<td>—</td>
<td>No</td>
<td>38</td>
</tr>
<tr>
<td>Vora et al (2020) 68</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>368</td>
<td>416</td>
<td>1</td>
<td>Unclear</td>
<td>No</td>
<td>56</td>
</tr>
<tr>
<td>Walts et al (2018) 58</td>
<td>United States 3</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>79 d</td>
<td>81 d</td>
<td>—</td>
<td>Both</td>
<td>No</td>
<td>33</td>
</tr>
<tr>
<td>Witt (2016) 79</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>32</td>
<td>32</td>
<td>&gt;1</td>
<td>—</td>
<td>No</td>
<td>63</td>
</tr>
<tr>
<td>Wu et al (2016) 60</td>
<td>United States 1</td>
<td>Prospective</td>
<td>AGEC</td>
<td>231</td>
<td>245</td>
<td>—</td>
<td>—</td>
<td>No</td>
<td>56</td>
</tr>
<tr>
<td>Yang et al (2016) 61</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>217</td>
<td>217</td>
<td>1</td>
<td>—</td>
<td>No</td>
<td>67</td>
</tr>
<tr>
<td>Zhang et al (2021) 69</td>
<td>United States 1</td>
<td>Retrospective</td>
<td>AGEC</td>
<td>120</td>
<td>127</td>
<td>1</td>
<td>Benign</td>
<td>No</td>
<td>44</td>
</tr>
</tbody>
</table>

aPercent of criteria (defined in Supplementary Appendix Table SA1) fulfilled, does not include observer variability criterium for studies with only one histopathologist (N=15). bFull-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. cResults separated by institution. dAGEC 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.
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, 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. 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 recategorization 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, 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, 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 sub-analysis. 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 utilized 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. 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. 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—Afirmà GEC, Afirmà 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


31. Schmarak F. bdpv: Inference and Design for Predictive Values in Diagnostic Tests. 2019. Available at: https://cran.r-project.org/package=bdpv


43. Chaudhary S, Hou Y, Shen R, et al. Impact of the Afirma gene expression classifier result on the surgical manage-


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


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