Supplementary Protocol

Molecular testing in oncocytic (Hürthle cell) Bethesda III/IV thyroid nodules: a systematic review and diagnostic test accuracy meta-analysis

Written protocol, version 1.0 (finalized 10 October 2026). This document is an English rendering of the written review protocol. It was finalized after a pilot, large-language-model-assisted screening round and a preliminary analysis, and before the second screening round and the final data extraction; the boundary-case rules in Section 4.2 were codified from the pilot round (see Section 10). The review was not registered in PROSPERO or any other public registry; the protocol is provided in full for transparency.

1. Objectives

In thyroid nodules with Bethesda category III/IV cytology that are oncocytic (Hürthle) cell-predominant, tested by molecular tests on fine-needle aspiration (FNA) material, the review asked:

1. What is the benign call rate (BCR)?
2. Against surgical histology (or follow-up), what are the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV)?
3. What is the risk of malignancy (ROM) by class of molecular alteration: no alteration; copy-number alteration (CNA) only; RAS-like; high-risk (TP53/TERT); BRAF V600E or kinase fusion?
4. Do estimates differ between platforms (Afirma GEC, Afirma GSC, ThyroSeq v2/v3, others)?

2. Eligibility criteria (PIRD framework)

Element	Definition
Population	FNA cytology Bethesda III (AUS/FLUS) or IV (FN/SFN), oncocytic/Hürthle cell-predominant nodules, or a separately reportable oncocytic subgroup within a broader indeterminate cohort. Accepted subgroup labels include AUS-H, AUS-HCT, AUS-O, AUS-PHC, III-O, FN-O, OFN, HCN, SHCN, FNHCT, SFNHCT, FNOF, "Hürthle cell predominant" and "oncocytic". Subgroups defined by histology (oncocytic/Hürthle adenoma or carcinoma) were also accepted, flagged "histology-defined", and used only in sensitivity analyses.
Index test	Any molecular test performed on FNA material: Afirma GEC; Afirma GSC (including Xpression Atlas); ThyroSeq v2/v3; ThyGeNEXT or ThyGenX with ThyraMIR; RosettaGX Reveal; other next-generation

	sequencing or mutation panels; copy-number (CNA) analysis; single-gene assays (e.g., BRAF V600E, HMGA2).
Reference standard	Surgical histopathology (preferred); clinical or sonographic follow-up; or, at minimum, reporting of the benign call rate.
Diagnostic measures	Extractable counts sufficient to derive at least one of: a 2 x 2 table (TP/FP/FN/TN); PPV; BCR; or malignancy rate by alteration class.

Inclusion required all of I1-I5:

5. I1 (population): Bethesda III/IV oncocytic-predominant nodules, or a separable oncocytic subgroup (as above).
6. I2 (index test): molecular testing on FNA material per the PIRD table.
7. I3 (reference standard): surgical histology, follow-up, or at least BCR.
8. I4 (extractable data): counts, or percentages with denominators, for at least one measure above.
9. I5 (design): original research with at least 5 oncocytic nodules in the subgroup.

Exclusion codes (first applicable code assigned):

Code	**Meaning**
E1	Not original research (narrative or systematic review, guideline, editorial, data-free letter, erratum).
E2	Case report, or case series with fewer than 5 oncocytic nodules.
E3	Population ineligible: no indeterminate thyroid FNA; Bethesda V/VI only; established carcinoma; other organs (e.g., parathyroid, renal or salivary oncocytic tumours).
E4	No separable oncocytic subgroup: term absent from the full text, or confined to the introduction or discussion, or appearing only as a histological diagnosis within pooled tables without test results.
E5	No molecular testing on FNA (tissue-only sequencing, plasma cfDNA, immunohistochemistry, imaging-only studies).
E7	Oncocytic subgroup mentioned but no extractable counts (no test-by-outcome numbers or denominators, and no BCR).
E8	Design ineligible (cost-effectiveness or decision models, analytical or bench validation, cell-line studies, usage-trend surveys without outcomes).
E10	Conference abstract only, with an existing full report (or insufficient data in the abstract).

E11	Full text unobtainable (including abstracts without retrievable text).

E9 (overlapping populations) was not adjudicated at screening; center and period were recorded for every included study (Section 6).

3. Information sources and search strategy

- Databases: PubMed/MEDLINE, Scopus, Web of Science Core Collection (SCIE and CPCI-S), and the Cochrane Library. Embase was not available institutionally and was not searched.
- Window: 1 January 2010 to the search date (8 October 2026); no language restrictions.
- Logic: thyroid terms AND (named tests OR (oncocytic/Hürthle terms AND molecular-testing terms)). Full strategies per database are given in Supplementary Table S1.
-
- Yield: 2774 records (PubMed 798, Scopus 1026, Web of Science 936, Cochrane 14); 1323 unique records after deduplication, each assigned a unique identifier (HC0001-HC1323).

4. Screening procedure

4.1 Title/abstract stage

- Only records obviously ineligible were excluded; any doubt was adjudicated MAYBE and passed to full text.
- Only codes E1, E2, E3, E5 and E8 were used at this stage; E4, E7 and E9 were not.
- Any original study of molecular testing on indeterminate thyroid FNA was retained (MAYBE) even if the abstract did not mention oncocytic cells, because subgroup data are frequently tabulated only in full text.
- Highly relevant reviews and meta-analyses were excluded (E1) but flagged for reference-list mining.

4.2 Full-text stage

Before judgment, every full text was searched with the case-insensitive expression "rthle|oncocyt|oxyphil" (robust to diacritics and broken encoding of "Hürthle"), and the context of every hit in the Results and tables, including supplementary tables where available, was read. Boundary rules applied consistently:

- A histological diagnosis of Hürthle adenoma or carcinoma listed only within pooled tables, without test-by-outcome data, was coded E4.
- Histology-defined subgroups with complete test-by-outcome counts were included, flagged "histology-defined", and used only in sensitivity analyses.
- Fragmentary histology-defined data were coded E7, except series restricted to TERT- or TP53-positive nodules that report oncocytic benign/malignant counts, which were included for alteration-class description only.
- Single-gene assays with an oncocytic subgroup of at least 5 nodules and histology were included, flagged "single-gene", and used for descriptive or sensitivity analyses only.

- Alteration-class data (e.g., "CNA only: x malignant / y benign") were included.
- Bethesda III-V cohorts not separable into III/IV were included, flagged, and used only in sensitivity analyses.
- Resected-only cohorts were included and flagged (spectrum bias).
- Conference abstracts with extractable data and no full report were included (sensitivity analyses); abstracts superseded by a full report were coded E10.
- Duplicate records of one study: one retained, the other flagged E9 with the retained identifier noted.
- Paediatric studies satisfying I1-I5 were included and flagged paediatric.
- The meaning of "AUS-O" was checked (2023 Bethesda "other atypia" vs oncocytic) before inclusion.
- Numbers derived by calculation, if any, were labeled as derived and never recorded as source values.

4.3 Dual review and agreement

Screening was performed in two rounds, both with the assistance of a large language model (Claude, Anthropic) applying this written protocol. The first round served as a pilot, and every inclusion and every uncertain decision was verified against the source by an investigator. The second round was performed by a separately instructed model session blinded to the first-round decisions. Rounds were merged by record identifier; agreement and Cohen's kappa were computed (title/abstract: INCLUDE+MAYBE vs EXCLUDE; full text: INCLUDE vs EXCLUDE, plus code agreement), and therefore describe the reproducibility of model-assisted screening rather than agreement between two human reviewers. All disagreements were resolved by an investigator re-reading the full text, with adjudication by a senior investigator, and reasons were documented.

5. Data extraction

- Study characteristics: country, center, period, design, cytological vs histological definition of the oncocytic subgroup, Bethesda category.
- Index test: test and version; positivity threshold used.
- Counts: nodules tested; valid results; benign (test-negative) calls; nodules resected; histological outcome by test result.
- Alteration class with the corresponding histology, where reported.
- A standardized form was used, and data were first extracted with large-language-model assistance; every value was traced to a specific table, figure or sentence. Every extracted value was verified against the source document by an investigator, and ambiguous or low-confidence values were re-checked by a second investigator. Authors were not contacted.

6. Overlapping populations (E9)

Where several publications from one center overlapped in period, one dataset per center and period entered each analysis, by priority: (1) most complete oncocytic-subgroup data; (2) largest; (3) most recent. A study could therefore contribute to one analysis but not another. Institutional

clusters monitored for overlap: UCLA; UPMC; Cleveland Clinic; University of Michigan; Ohio State; University of Pennsylvania; University of Miami; BIDMC/BWH/MGH; Yale.

7. Outcome definitions

- Positive index test: a "suspicious" Afirma result; for ThyroSeq, the official positive call where available, or a result reported as positive by the study (risk of malignancy above 10%, or intermediate/high probability of cancer). Two included studies defined any reportable alteration as positive and were excluded in a sensitivity analysis.
- Malignancy: all carcinomas, including oncocytic carcinoma, papillary carcinoma and its variants, follicular carcinoma and medullary carcinoma. NIFTP was counted as benign in the primary analysis where separable, and as malignant in a sensitivity analysis; where a study pooled NIFTP with malignant outcomes, the published classification was retained and the study examined in sensitivity analyses.
- Scenario A (resected only): 2 x 2 tables built from nodules with surgical histology; specificity is underestimated because test-negative nodules are under-represented.
- Scenario B (intention-to-test): unoperated test-negative nodules counted as true negatives; specificity and NPV are overestimated. The true value lies between the scenarios; both are reported.
- BCR: test-negative results divided by nodules with a valid result. PPV: malignancy among test-positive resected nodules. Alteration-class ROM: malignancy among resected nodules within each class.
- Alteration classes: (1) no alteration; (2) CNA only; (3) RAS-like (RAS, EIF1AX, PAX8::PPARG, BRAF K601E or THADA fusion, with or without CNA); (4) high-risk (TP53 and/or TERT promoter); (5) BRAF V600E or kinase fusions (RET/NTRK/ALK).

8. Risk of bias and certainty of evidence

Risk of bias and applicability were assessed with QUADAS-2, tailored with signaling questions on consecutive or complete enrolment of oncocytic nodules; prespecified test thresholds; blinding of pathologists to molecular results; partial and differential verification; and exclusion of nondiagnostic or unresected nodules. Judgments were drafted with large-language-model assistance, each with a written justification, and were reviewed and finalized by two investigators. Certainty of evidence was rated with the GRADE approach for diagnostic test accuracy.

9. Statistical analysis plan

- Single proportions (BCR, PPV, NPV, ROM by class): univariate binomial-normal random-effects GLMM (logit link), Gauss–Hermite quadrature (40 nodes); standard errors from the full observed information matrix; exact (Clopper–Pearson) intervals when all studies show 0% or 100%.
- Sensitivity and specificity: bivariate binomial-normal GLMM (Chu and Cole); summary operating points in SROC space when k >= 4; univariate fallbacks where the bivariate model is not estimable (fewer than 4 datasets, or all studies with zero false negatives).

- Platform comparisons: Wald tests of differences between pooled logits; meta-regression as a follow-up.
- CNA-only vs RAS-like ROM: Fisher's exact test on pooled counts.
- Planned sensitivity analyses: (i) excluding resected-only cohorts; (ii) adding the high-volume-center conference abstract; (iii) adding histology-defined cohorts; (iv) NIFTP counted as malignant (equivalently, excluding studies pooling NIFTP with malignancy); (v) leave-one-out; (vi) ThyroSeq v3 restricted to the official classifier call; (vii) Bethesda III vs IV stratification; (viii) stratification by cytology nomenclature edition (2017 vs 2023 Bethesda).
- Small-study effects: Deeks' funnel plot asymmetry test when k >= 10, with test platform as a covariate.
- Verification bias treated qualitatively (Begg-Greenes framework), bracketed by scenarios A and B.

10. Protocol history and deviations

Version 1.0 (finalized 10 October 2026) was written after the pilot, AI-assisted first screening round and a preliminary analysis of the first-round data. The boundary-case rules in Section 4.2 were codified from that pilot round and fixed in writing before the blinded second screening round; the eligibility criteria, outcome definitions, verification scenarios, overlap rule and primary analyses were unchanged thereafter, and no changes to eligibility, outcomes or primary analyses were made after the final data extraction. Deviations from the plan: hierarchical SROC curves were not drawn because between-study variance was poorly estimated at small k (summary operating points with confidence regions are shown instead); sensitivity analysis (viii), stratification by nomenclature edition, was not performed; the planned 40-node Gauss–Hermite quadrature for univariate models was increased to 150 nodes (with 50 × 50 nodes for bivariate models) for numerical stability; and GRADE Summary-of-Findings rows were added at the writing stage as required by PRISMA-DTA. The planned sensitivity analysis counting NIFTP as malignant (Section 9, item iv) was replaced by excluding, in sensitivity analyses, the studies that pooled NIFTP with malignancy from each primary analysis in which they appeared (GSC scenario A; GSC and GEC scenario B; the RAS-like class; and the GEC scenario A analysis, from which all four pre-reclassification studies were excluded). In the histology-defined sensitivity analysis each patient was counted once (Patel 2018 preferred to Alexander 2012; the GEC arm of Walts 2018 preferred to its RosettaGX Reveal arm), and Onken 2021 was excluded from the GEC accuracy analyses because its cohort overlapped Brauner 2015, in keeping with the one-dataset-per-center-and-period rule. Subgroups defined by the presence (rather than predominance) of oncocytic cells were accepted where results for indeterminate cytology were reported separately (Attia 2026, Endo 2019, Khan 2026), a relaxation of criterion I1; excluding Attia 2026 and Endo 2019 raised the pooled GSC benign call rate from 74.2% to 75.4% (Supplementary Table S6). Sensitivity analysis (vi), restriction of ThyroSeq v3 to the official classifier call, was replaced by excluding the two studies that defined any reportable alteration as positive, because all other studies already used the official call or a study-specific positivity threshold. The planned meta-regression of accuracy on platform was not performed: platform was largely confounded with center and period, and fewer than ten datasets per platform were available. For Attia 2026, the published Hürthle-cell-positive table was used as reported; that

table counted unoperated GSC-suspicious nodules as benign (false positives), and excluding Attia 2026 was examined in a sensitivity analysis. Khan 2026 (conference abstract) reported percentages only; its 2×2 cells were reconstructed from the reported denominators (total 150, versus 153 resected stated in the abstract, a rounding discrepancy). Keutgen 2012 (four-miRNA panel) contributed to descriptive synthesis only, with counts reconstructed as noted in Supplementary Table S2.