



# A COMPREHENSIVE SYSTEMATIC REVIEW AND META-ANALYSIS: THE FUTURE OF ARTIFICIAL INTELLIGENCE IN DIAGNOSTIC OTOTOLOGY — OPPORTUNITIES AND BARRIERS TO CLINICAL IMPLEMENTATION

Andreyne Lidya Zhelin Tuuk <sup>1\*</sup>, Hanski R. P. Legrans <sup>2</sup>

<sup>1</sup> General Practitioner, RSUD Kota Kotamobagu, Kotamobagu, North Sulawesi,  
Indonesia

<sup>2</sup> Otolaryngologist, RSUD Kota Kotamobagu, Kotamobagu, North Sulawesi,  
Indonesia

**\*Corresponding author:** Andreyne Lidya Zhelin Tuuk., MD, RSUD Kota  
Kotamobagu, Kotamobagu, North Sulawesi, Indonesia. E-mail:  
[Andreynetuuk@gmail.com](mailto:Andreynetuuk@gmail.com)

## Article History :

Received date : 2026/07/12  
Revised date : 2026/08/05  
Accepted date : 2026/09/23  
Published date : 2026/10/16



**Copyright:** © 2024 by the  
authors. Submitted for possible  
open access publication under  
the terms and conditions of the  
Creative Commons Attribution  
(BY NC) license  
(<https://creativecommons.org/licenses/by-nc/4.0/>).

E-ISSN :

ISSN 3048-1368



P-ISSN

ISSN 3048-1376



## ABSTRACT

**Introduction:** Diagnostic otology relies on operator-dependent interpretation of otoscopy, temporal bone computed tomography (CT), audiometry and vestibular testing, and diagnostic accuracy outside specialist settings remains suboptimal. Artificial intelligence (AI) has been proposed as a scalable decision-support solution, but its pooled accuracy, robustness beyond development data, comparative value against clinicians and barriers to clinical implementation have not been synthesised across all diagnostic otology modalities. This review aimed to quantify the diagnostic performance of AI across otoscopic, radiological, audiological and vestibular applications and to map opportunities and barriers to implementation.

**Methods:** The study strictly adhered to the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) 2020 guidelines. Randomised controlled trials, cohort, case-control and cross-sectional diagnostic studies (prospective or retrospective) evaluating an AI index test against a clinical reference standard were eligible; systematic reviews were excluded from synthesis. Risk of bias was appraised with QUADAS-2 supplemented by QUADAS-AI considerations. Accuracy, sensitivity and specificity were pooled on the logit scale using random-effects models (restricted maximum likelihood with modified Hartung-Knapp-Sidik-Jonkman confidence intervals) with 95% prediction intervals (PI); risk differences (RD) quantified AI-clinician comparisons. Subgroup analyses, meta-regression, influence diagnostics, small-study tests and GRADE certainty assessment were performed.

**Results:** Of 139 unique records, 64 primary studies (2019–2026; otoscopy 37, temporal bone CT 14, audiology 6, vestibular 7) were included. No randomised trial of an AI diagnostic tool in otology was identified. Pooled internal-validation accuracy was 92.5% (95% CI 90.0–94.5) ( $k = 36$ ; PI 64.7–98.8%;  $I^2 = 97.8\%$ ,  $\tau^2 = 0.858$ ): otoscopy 93.2%, temporal bone CT 90.2%, audiometry/tympanometry 97.6% and vestibular testing 84.0% (between-modality  $P = 0.003$ ). Pooled sensitivity and specificity were 92.6% (95% CI 89.4–94.8) and 89.5% (95% CI 74.3–96.1), respectively (diagnostic odds ratio 105; summary AUROC 0.957). Accuracy fell to 81.2% (95% CI 59.1–92.8) in external, prospective or point-of-care cohorts ( $P = 0.026$  versus internal validation), and studies using surgical, histopathological or objective reference standards reported lower accuracy than those using expert annotation (89.0% vs 93.6%;  $P = 0.041$ ). Task-

specific AI outperformed clinicians on identical cases (RD +18.2 percentage points, 95% CI 7.3 to 29.0), and AI assistance increased clinician accuracy (RD +9.4 points, 95% CI 1.4 to 17.4), whereas general-purpose multimodal large language models underperformed specialists (RD -25.0 points). Of 47 meta-analysed studies, 34 (72.3%) had at least one high risk-of-bias domain; certainty of evidence was low to very low.

**Discussion:** Task-specific AI achieves high diagnostic accuracy across diagnostic otology and consistently exceeds non-specialist performance, supporting its role as a triage and decision-support adjunct, particularly in primary care, tele-otology and resource-limited settings. However, extreme heterogeneity, wide prediction intervals, performance attenuation under domain shift, spectrum and reference-standard bias, scarce prospective evidence and the absence of outcome trials constitute the principal barriers to implementation.

**Conclusion:** AI is a promising but not yet implementation-ready component of diagnostic otology. Multicentre prospective validation with surgical or objective reference standards, standardised image acquisition, transparent reporting (STARD-AI/TRIPOD+AI-aligned) and pragmatic randomised trials measuring patient-important outcomes are required before routine deployment.

**Keywords:** artificial intelligence; deep learning; otoscopy; otitis media; cholesteatoma; temporal bone computed tomography; audiometry; nystagmus; diagnostic accuracy; meta-analysis

---

## INTRODUCTION

---

### Background

Hearing loss is among the leading causes of years lived with disability worldwide. The Global Burden of Disease 2019 analysis estimated that approximately 1.57 billion people—one in five—lived with hearing loss, a figure projected to reach 2.45 billion by 2050, with the heaviest burden concentrated in countries with limited health-care access and quality. The global economic cost of unaddressed hearing loss has been estimated to exceed 981 billion international dollars per year.<sup>1,2</sup>

Otitis media remains the leading preventable cause of hearing loss in childhood: in the Global Burden of Disease 2021 study, 46.9% of hearing loss in children and adolescents, and 68.7% in children younger than five years, was attributed to otitis media. Accurate and timely otological diagnosis is therefore central to preventing language delay, educational disadvantage and complications such as cholesteatoma, intratemporal and intracranial sequelae, and permanent sensorineural damage.<sup>1,3</sup>

Diagnostic otology is nevertheless intrinsically operator-dependent. Otoscopic recognition of tympanic membrane bulging, effusion, retraction or perforation requires training, adequate visualisation and frequently pneumatic otoscopy, and diagnostic agreement among primary care clinicians is consistently modest; experienced paediatricians viewing standardised otoscopic videos have achieved accuracy of approximately 70% for acute otitis media (AOM) and around 45% for otitis media with effusion (OME). Misclassification drives inappropriate antibiotic prescribing, missed effusions with conductive hearing loss and delayed referral.<sup>4</sup>

Beyond otoscopy, interpretation of high-resolution temporal bone CT for chronic otitis media (COM) and cholesteatoma, classification of pure-tone audiograms and tympanograms, and recognition of positional nystagmus in benign paroxysmal positional vertigo (BPPV) all require subspecialty expertise that is scarce in rural, remote and low- and middle-income settings, including Indonesia. Digital otoscopes, smartphone-attached endoscopes and video-oculography have expanded image capture, but randomised trials of digital otoscopy alone have shown only modest or inconsistent improvements in diagnostic agreement and antibiotic prescribing, indicating that image acquisition without interpretive support is insufficient.<sup>5–8</sup>

Artificial intelligence (AI), particularly deep learning with convolutional neural networks (CNNs), recurrent and transformer architectures and, more recently, multimodal large language models (LLMs), can learn discriminative representations directly from images, signals and videos. Across medical imaging, deep learning has achieved diagnostic performance comparable

with health-care professionals under laboratory conditions, yet with consistently poor reporting, limited external validation and high heterogeneity. Diagnostic otology—where the input data (otoscopic images, CT slices, audiograms, eye-movement traces) are standardised, digital and abundant—is a natural candidate for AI-enabled triage and decision support.<sup>9–11</sup>

### **Research gap**

Previous evidence syntheses have focused almost exclusively on otoscopic image classification. Habib and colleagues pooled AI accuracy for otoscopy, Cao and colleagues reported pooled sensitivity and specificity for tympanic membrane images, and Song and colleagues narratively reviewed image-based AI for middle ear disease. These reviews predate most temporal bone CT, audiological, vestibular and LLM studies, did not jointly quantify performance under external validation, did not pool head-to-head AI–clinician or AI-assisted comparisons as effect sizes, and did not systematically map translational barriers. A separate review addressed machine learning for BPPV but did not synthesise accuracy quantitatively.<sup>12–15</sup>

Concurrently, empirical evidence from radiology shows that most deep learning algorithms lose performance on external data, and evidence from otoscopy indicates substantial internal-to-external attenuation and dataset-specific artefacts. Whether the high accuracies reported in development studies translate into clinically dependable performance in diagnostic otology—and which methodological and contextual factors explain the extreme between-study heterogeneity—therefore remains unresolved.<sup>16–18</sup>

### **Novelty**

To our knowledge, this is the first systematic review and meta-analysis to (i) integrate all four diagnostic otology modalities—otoscopy/otoendoscopy, temporal bone CT, audiometry/tympanometry and vestibular testing—within a single analytical framework; (ii) quantify the internal-to-external validation gap using pooled external cohorts; (iii) pool head-to-head AI-versus-clinician and AI-assisted-versus-unassisted comparisons as risk differences; (iv) contrast task-specific deep learning with general-purpose multimodal LLMs; (v) test reference-standard, task-complexity and design effects through subgroup analysis and meta-regression; and (vi) translate the quantitative findings into an evidence-based opportunities–barriers framework with GRADE certainty ratings.<sup>19,20</sup>

### **Objectives**

The primary objective was to estimate the pooled diagnostic accuracy of AI systems for otological diagnosis across modalities. Secondary objectives were to estimate pooled sensitivity, specificity, diagnostic odds ratio and area under the receiver operating characteristic curve (AUROC); to compare AI with clinicians and quantify the incremental value of AI assistance; to

assess performance under external, prospective and point-of-care validation; to evaluate general-purpose multimodal LLMs; to explore sources of heterogeneity; and to synthesise opportunities and barriers to clinical implementation.<sup>19</sup>

### **Hypotheses**

We hypothesised that (H1) task-specific AI achieves a pooled diagnostic accuracy of at least 90% under internal validation across diagnostic otology modalities; (H2) performance is significantly lower in external, prospective or point-of-care validation than in internal validation; (H3) AI is more accurate than clinicians when evaluated on identical cases and improves clinician accuracy when used as an assistant; and (H4) studies using surgical, histopathological or objective reference standards report lower accuracy than studies relying on expert image annotation, reflecting spectrum and reference-standard bias.<sup>9,16</sup>

### **Significance and benefits**

For clinicians, the review clarifies where AI can be trusted as a triage or second-reader tool and where caution is warranted. For health systems—particularly in archipelagic and resource-limited settings such as Indonesia, where otolaryngologists are concentrated in urban centres—it informs the potential role of smartphone-based and tele-otology AI in primary care referral pathways. For researchers, developers and regulators, it identifies methodological priorities, including external validation, surgical reference standards, prospective deployment studies and randomised evaluation of patient-important outcomes.<sup>1,21,22</sup>

---

## **METHODS**

---

### **Protocol, reporting standard and registration**

This systematic review and meta-analysis was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement and its explanation and elaboration document, with diagnostic-accuracy-specific items considered throughout. Risk-of-bias appraisal followed QUADAS-2 informed by the AI-specific considerations proposed for QUADAS-AI and the recently published QUADAS-3 framework. The protocol was not prospectively registered; all eligibility criteria, outcomes and analytical decisions were specified before data synthesis.<sup>19,20,23–25</sup>

### **Eligibility criteria**

Eligibility was defined using a PICOTS framework. Studies were eligible if they (i) evaluated an AI index test (machine learning, deep learning, automated machine learning or multimodal LLM) for the diagnosis, classification or triage of an otological condition from otoscopic/otoendoscopic images or videos, temporal bone CT, pure-tone audiometry, tympanometry or vestibular eye-movement recordings; (ii) used a clinical reference standard

(expert otological diagnosis, consensus labelling, intraoperative or histopathological findings, audiometric thresholds or a clinical vestibular diagnosis); (iii) reported, or allowed derivation of, accuracy, sensitivity, specificity or AUROC with an identifiable test denominator; and (iv) used a randomised controlled, cohort (prospective or retrospective), case–control, cross-sectional or other observational diagnostic design. Studies were restricted to the last 10–15 years (January 2014 onward). Systematic reviews and meta-analyses were not included in the characteristics tables or quantitative synthesis but were retained for the Discussion.<sup>19,20</sup>

**Table 1. Eligibility criteria (PICOTS framework)**

| Domain                          | Inclusion criteria                                                                                                                                                                                                                 | Exclusion criteria                                                                                                                                               |
|---------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Population                      | Patients of any age undergoing otological assessment (otoscopy/otoendoscopy, temporal bone CT, pure-tone audiometry, tympanometry, video-oculography/videonystagmography)                                                          | Animal or veterinary datasets; phantom or purely synthetic data without clinical test set                                                                        |
| Index test                      | AI system (machine learning, deep learning, AutoML, ensemble, hybrid, sequence models, multimodal large language models) producing a diagnostic, classification or triage output                                                   | Segmentation, digitisation or image retrieval without patient/image-level diagnostic output; hardware-coupled devices in which the AI component is not separable |
| Comparator / reference standard | Expert otological diagnosis or consensus labels; intraoperative, myringotomy or histopathological findings; audiometric air–bone gap; clinical vestibular diagnosis. Optional human comparator (clinicians reading the same cases) | Reference standard derived from non-clinical group membership or simulated cases                                                                                 |
| Outcomes                        | Accuracy; sensitivity; specificity; AUROC; diagnostic odds ratio; AI–clinician accuracy difference; change in clinician accuracy with AI assistance                                                                                | No extractable or internally consistent test-set denominator                                                                                                     |
| Timing                          | Published January 2014 to 14 October                                                                                                                                                                                               | Before 2014                                                                                                                                                      |

|              |                                                                                                                                                      |                                                                                                                                                                                                                                    |
|--------------|------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|              | 2026 (last 10–15 years)                                                                                                                              |                                                                                                                                                                                                                                    |
| Study design | Randomised controlled trials; prospective or retrospective cohort; case–control; cross-sectional and other observational diagnostic-accuracy studies | Systematic reviews/meta-analyses (Discussion only); narrative reviews; editorials; conference abstracts/engineering proceedings without full peer-reviewed clinical reporting; overlapping cohorts (most complete report retained) |

### Search Strategy

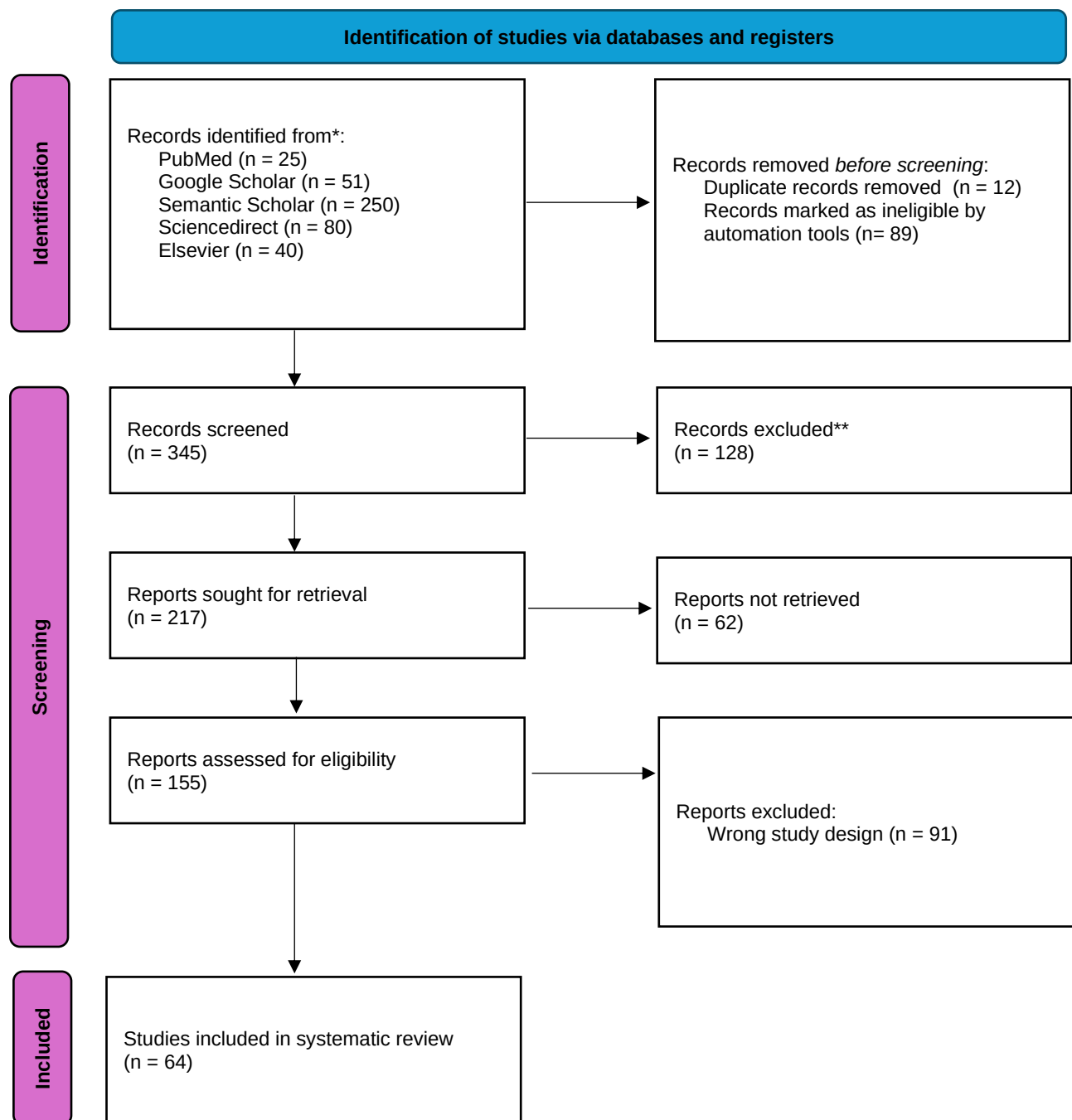
The keywords used for this research based PICO :

| PICO Element            | Keyword 1               | Keyword 2                    | Keyword 3               |
|-------------------------|-------------------------|------------------------------|-------------------------|
| <b>P</b> (Population)   | Diagnostic Otology      | Ear Diseases                 | Otologic Patients       |
| <b>I</b> (Intervention) | Artificial Intelligence | Deep Learning                | Clinical Implementation |
| <b>C</b> (Comparison)   | Clinician Diagnosis     | Otolaryngologist Performance | Reference Standard      |
| <b>O</b> (Outcome)      | Diagnostic Accuracy     | Sensitivity and Specificity  | Implementation Barriers |

The Boolean MeSH keywords inputted on databases for this research are: ("Diagnostic Otology" OR "Ear Diseases" OR "Otologic Patients") AND ("Artificial Intelligence" OR "Deep Learning" OR "Clinical Implementation") AND ("Clinician Diagnosis" OR "Otolaryngologist Performance" OR "Reference Standard") AND ("Diagnostic Accuracy" OR "Sensitivity and Specificity" OR "Implementation Barriers")

## Article Search Strategy

| Database         | Keywords                                                                                                                                                                                                                                                                                                                               | Hits |
|------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| Pubmed           | <i>("Diagnostic Otology" OR "Ear Diseases" OR "Otologic Patients") AND ("Artificial Intelligence" OR "Deep Learning" OR "Clinical Implementation") AND ("Clinician Diagnosis" OR "Otolaryngologist Performance" OR "Reference Standard") AND ("Diagnostic Accuracy" OR "Sensitivity and Specificity" OR "Implementation Barriers")</i> | 25   |
| Google Scholar   | <i>("Diagnostic Otology" OR "Ear Diseases" OR "Otologic Patients") AND ("Artificial Intelligence" OR "Deep Learning" OR "Clinical Implementation") AND ("Clinician Diagnosis" OR "Otolaryngologist Performance" OR "Reference Standard") AND ("Diagnostic Accuracy" OR "Sensitivity and Specificity" OR "Implementation Barriers")</i> | 51   |
| Semantic Scholar | <i>("Diagnostic Otology" OR "Ear Diseases" OR "Otologic Patients") AND ("Artificial Intelligence" OR "Deep Learning" OR "Clinical Implementation") AND ("Clinician Diagnosis" OR "Otolaryngologist Performance" OR "Reference Standard") AND ("Diagnostic Accuracy" OR "Sensitivity and Specificity" OR "Implementation Barriers")</i> | 250  |
| Science-direct   | <i>("Diagnostic Otology" OR "Ear Diseases" OR "Otologic Patients") AND ("Artificial Intelligence" OR "Deep Learning" OR "Clinical Implementation") AND ("Diagnostic Accuracy" OR "Sensitivity and Specificity" OR "Implementation Barriers")</i>                                                                                       | 80   |
| Cochrane         | <i>("Diagnostic Otology" OR "Ear Diseases" OR "Otologic Patients") AND ("Artificial Intelligence" OR "Deep Learning" OR "Clinical Implementation") AND ("Diagnostic Accuracy" OR "Sensitivity and Specificity" OR "Implementation Barriers")</i>                                                                                       | 40   |



**Article search flowchart**

### Study selection

Records were de-duplicated across queries and screened on title and abstract against the eligibility criteria, followed by full-text assessment. Where several reports described the same or overlapping cohorts, the report with the most complete diagnostic accuracy data was retained. Reasons for full-text exclusion were recorded, and the selection process is summarised in a PRISMA 2020 flow diagram.<sup>19,23</sup>

## Data extraction

A piloted extraction form captured bibliographic data, country, design and setting, population (paediatric, adult or mixed), modality, target conditions and number of diagnostic classes, dataset size and partitioning, AI architecture and training strategy, validation approach (hold-out, cross-validation, external, prospective or point-of-care), reference standard, human comparators and diagnostic metrics. For accuracy synthesis, the number of correctly classified test units and the test denominator were extracted; when only percentages were reported, counts were back-calculated from the reported test denominator. When the denominator had to be approximated from the described data split (Chen YC 2022; Guo 2025), reconstructed from sensitivity and specificity (Surapaneni 2026), inferred from reported proportions (Suresh 2024) or derived from confidence intervals (Crowson 2023), this was flagged and tested in sensitivity analyses. For multi-model studies, the best-performing model prespecified by the authors as the primary model was extracted. Sensitivity and specificity were extracted with the disease-positive class as the reference category; where authors reported the normal class as positive (Dubois 2024), counts were re-oriented.<sup>22,26–28</sup>

## Risk of bias and applicability

Risk of bias and applicability concerns were assessed using QUADAS-2 across four domains—patient selection, index test, reference standard, and flow and timing—with AI-specific signalling considerations derived from QUADAS-AI: curated or web-sourced image libraries and case–control enrichment (patient selection); absence of a locked independent test set, model selection on the evaluation data, image-level partitioning with risk of patient-level leakage and data augmentation before partitioning (index test); single-reader or unverified labels versus surgical, myringotomy or histopathological confirmation (reference standard); and exclusion of non-diagnostic images or image/clip-level analysis of clustered data (flow and timing). Each domain was rated as low, high or unclear risk.<sup>20,24,25,29</sup>

## Risk of bias and applicability

Patient selection was the most problematic domain: 26 studies (55.3%) were at high risk because they used curated institutional image libraries, public or web-derived datasets, representative-case enrichment or case–control designs, and only 11 (23.4%) enrolled consecutive or prospectively sampled patients. Fifteen studies (31.9%) were at high risk in the index-test domain, chiefly because performance was estimated by cross-validation or on a validation set also used for model selection, or because image-level partitioning or pre-partition augmentation risked patient-level leakage. The reference standard was unclear in 24 studies (51.1%) that relied on single-reader or unspecified expert labels, whereas 21 used surgical,

myringotomy, histopathological, audiometric or rigorous consensus references. Flow and timing were unclear in 23 studies, and three analysed clustered images or clips without accounting for patient-level dependence. Overall, 34 studies (72.3%) had at least one domain at high risk and only five (10.6%)—Prakairunghong 2026, Chen BJ 2024, Yağcı 2026, Anh 2022 and Wu P 2023—were at low risk across all four domains. Applicability concerns were limited, arising mainly from operative-only paediatric cohorts and web-sourced images.<sup>20,29,41,43,45,64,91</sup>

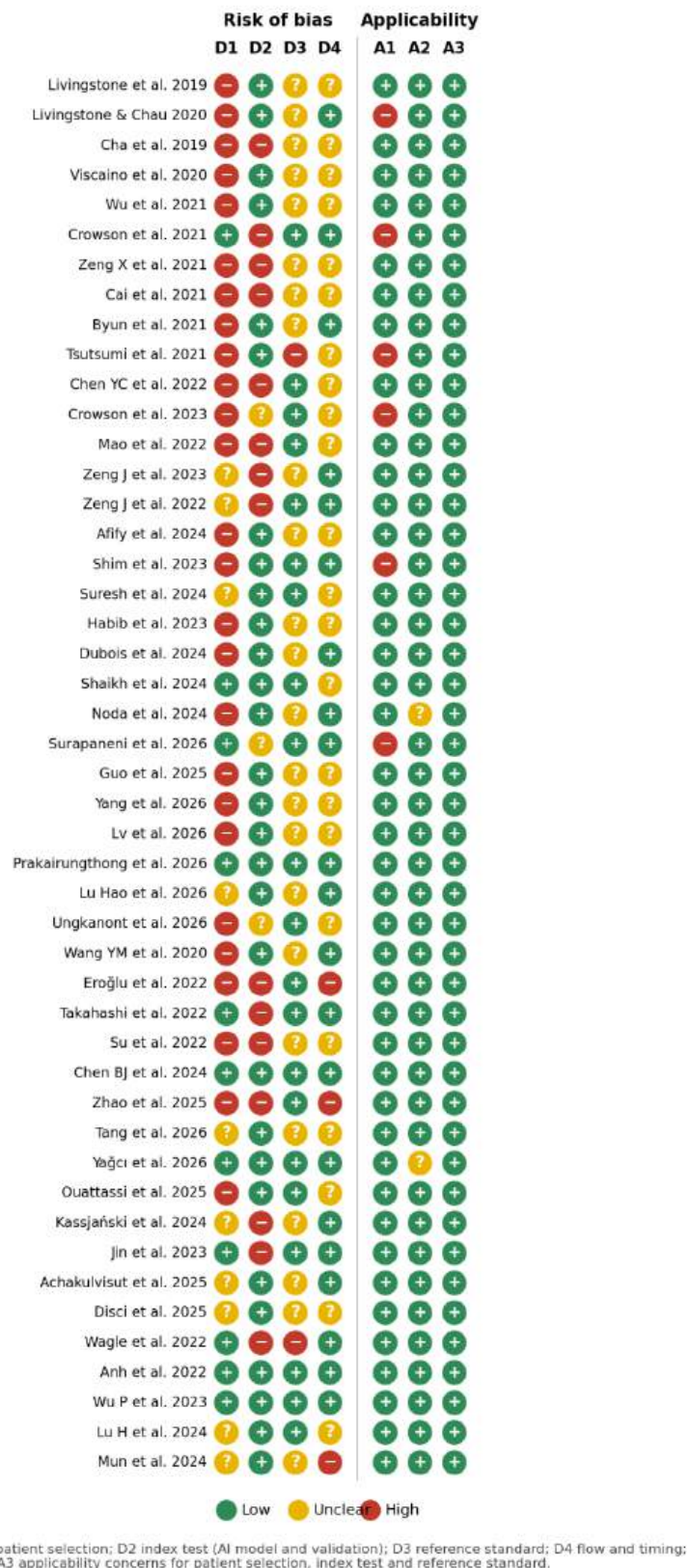


Figure 1. QUADAS-2 traffic-light plot (with QUADAS-AI signalling considerations) for the 47 meta-analysed studies. Green (+) low, yellow (?) unclear, red (−) high risk or concern.

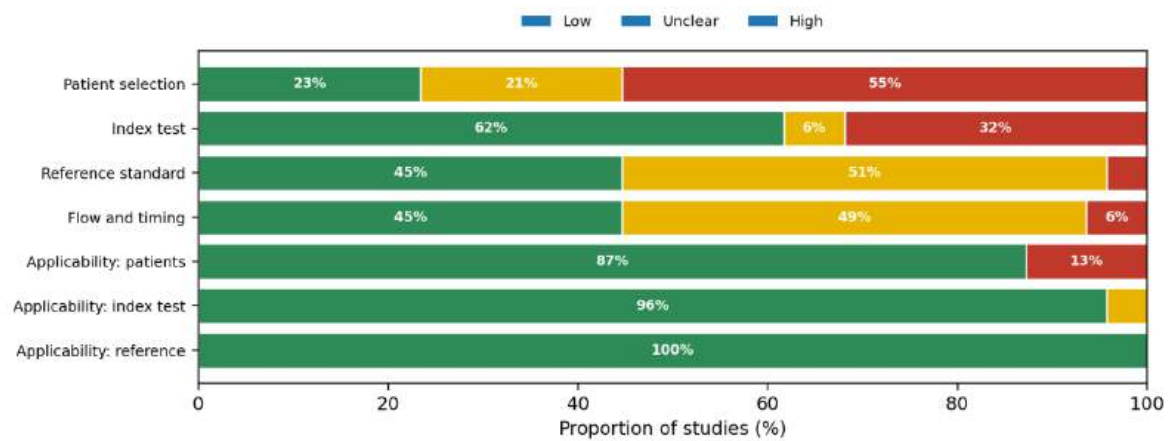


Figure 2. Summary of risk-of-bias and applicability judgements across QUADAS-2 domains (proportion of studies).

**Table 2. Domain-level QUADAS-2 judgements and principal rationale (meta-analysed studies, n = 47)**

| Study                   | Patient selection | Index test | Reference standard | Flow & timing | Principal rationale                                                                             |
|-------------------------|-------------------|------------|--------------------|---------------|-------------------------------------------------------------------------------------------------|
| Livingstone et al. 2019 | High              | Low        | Unclear            | Unclear       | Curated institutional image library; 20% held-out validation (45 images); single-reader labels  |
| Livingstone & Chau 2020 | High              | Low        | Unclear            | Low           | Images partly sourced from Google Images/open repositories; 89-image hold-out; consensus labels |
| Cha et al. 2019         | High              | High       | Unclear            | Unclear       | Retrospective otoendoscopic database; 5-fold cross-validation without locked test set           |
| Viscaino et al. 2020    | High              | Low        | Unclear            | Unclear       | Selected images of four conditions; separate 160-image test set                                 |
| Wu et al. 2021          | High              | Low        | Unclear            | Unclear       | Institutional images with quality filtering; 1,500-image hold-out plus 102 prospective          |

|                      |         |         |         |         |                                                                                                    |
|----------------------|---------|---------|---------|---------|----------------------------------------------------------------------------------------------------|
|                      |         |         |         |         | smartphone images                                                                                  |
| Crowson et al. 2021  | Low     | High    | Low     | Low     | Intraoperative images of children listed for tympanostomy; cross-validation; myringotomy reference |
| Zeng X et al. 2021   | High    | High    | Unclear | Unclear | 80/20 split with model selection on validation set; no independent test set                        |
| Cai et al. 2021      | High    | High    | Unclear | Unclear | Three-fold cross-validation on 6,066 images; specialist labels                                     |
| Byun et al. 2021     | High    | Low     | Unclear | Low     | 71 representative cases selected for testing (enriched spectrum)                                   |
| Tsutsumi et al. 2021 | High    | Low     | High    | Unclear | Publicly available online images with unverified provenance of labels                              |
| Chen YC et al. 2022  | High    | High    | Low     | Unclear | 8:2 random split; three-specialist consensus labels                                                |
| Crowson et al. 2023  | High    | Unclear | Low     | Unclear | Operative cohort; test partition not fully described; intraoperative reference                     |
| Mao et al. 2022      | High    | High    | Low     | Unclear | Validation set used for model selection; mycological confirmation                                  |
| Zeng J et al. 2023   | Unclear | High    | Unclear | Low     | Multicentre OME images; three-fold cross-validation                                                |
| Zeng J et al. 2022   | Unclear | High    | Low     | Low     | 80/20 split; audiometric air-bone gap as objective reference                                       |
| Afify et al. 2024    | High    | Low     | Unclear | Unclear | Public Chilean otoscopy dataset; separate 264-image test set                                       |
| Shim et al. 2023     | High    | Low     | Low     | Low     | Poor-quality images                                                                                |

|                             |         |         |         |         |                                                                                  |
|-----------------------------|---------|---------|---------|---------|----------------------------------------------------------------------------------|
|                             |         |         |         |         | excluded; OME confirmed by myringotomy; independent 100-image set                |
| Suresh et al. 2024          | Unclear | Low     | Low     | Unclear | Point-of-care comparison in operative setting; brief report                      |
| Habib et al. 2023           | High    | Low     | Unclear | Unclear | Three public cohorts; explicit external (cross-cohort) testing                   |
| Dubois et al. 2024          | High    | Low     | Unclear | Low     | Class balancing and near-duplicate images in development data; held-out test set |
| Shaikh et al. 2024          | Low     | Low     | Low     | Unclear | Consecutive sick/well visits; validated otoscopists; held-out test set           |
| Noda et al. 2024            | High    | Low     | Unclear | Low     | Selected images; zero-shot LLM with prompt optimised on separate images          |
| Surapaneni et al. 2026      | Low     | Unclear | Low     | Low     | Multicentre operative cohort; partition not reported; myringotomy reference      |
| Guo et al. 2025             | High    | Low     | Unclear | Unclear | 60/20/20 split; single-centre labels                                             |
| Yang et al. 2026            | High    | Low     | Unclear | Unclear | Independent 122-image test set; retrospective selection                          |
| Lv et al. 2026              | High    | Low     | Unclear | Unclear | Balanced locked test set (200/class) enriches spectrum                           |
| Prakairungthong et al. 2026 | Low     | Low     | Low     | Low     | Prospective external validation; three-otolaryngologist consensus                |
| Lu Hao et al. 2026          | Unclear | Low     | Unclear | Low     | External paired video-tympanometry cohort                                        |
| Ungkanont et al. 2026       | High    | Unclear | Low     | Unclear | Small paediatric dataset; surgical findings as reference                         |
| Wang YM et al.              | High    | Low     | Unclear | Low     | COM patients; 15% hold-out;                                                      |

|                          |         |      |         |         |                                                                                             |
|--------------------------|---------|------|---------|---------|---------------------------------------------------------------------------------------------|
| 2020                     |         |      |         |         | derivation of ground-truth labels not detailed in the abstract; comparison with six experts |
| Eroğlu et al. 2022       | High    | High | Low     | High    | Case-control (surgical COM vs controls); image-level split risks patient leakage            |
| Takahashi et al. 2022    | Low     | High | Low     | Low     | Consecutive surgical cholesteatoma; eight-fold cross-validation                             |
| Su et al. 2022           | High    | High | Unclear | Unclear | Noise augmentation before partitioning; potential leakage                                   |
| Chen BJ et al. 2024      | Low     | Low  | Low     | Low     | Two institutes; external and prospective validation; surgical reference                     |
| Zhao et al. 2025         | High    | High | Low     | High    | Case-control image sets; five-fold CV; image-level analysis                                 |
| Tang et al. 2026         | Unclear | Low  | Unclear | Unclear | Separate 406-image test set; reader study                                                   |
| Yağcı et al. 2026        | Low     | Low  | Low     | Low     | Consecutive surgical patients; histopathology; zero-shot LLM                                |
| Ouattassi et al. 2025    | High    | Low  | Low     | Unclear | Case-control design; external 125-case validation                                           |
| Kassjański et al. 2024   | Unclear | High | Unclear | Low     | Ten-fold CV on 15,046 audiograms; audiologist labels                                        |
| Jin et al. 2023          | Low     | High | Low     | Low     | Screening-trial cohorts; ten-fold CV; audiologist reference                                 |
| Achakulvisut et al. 2025 | Unclear | Low  | Unclear | Low     | 70/30 split; otolaryngologist interpretation as reference                                   |
| Disci et al. 2025        | Unclear | Low  | Unclear | Unclear | 800/200 split; single-centre audiology data                                                 |
| Wagle et al.             | Low     | High | High    | Low     | AVERT trial videos; cross-                                                                  |

|                  |         |     |         |         |                                                                |
|------------------|---------|-----|---------|---------|----------------------------------------------------------------|
| 2022             |         |     |         |         | validation; single-expert labels                               |
| Anh et al. 2022  | Low     | Low | Low     | Low     | Consecutive neuro-otological work-up; 25% test set             |
| Wu P et al. 2023 | Low     | Low | Low     | Low     | Separate test cohort (n=703) for generalisability              |
| Lu H et al. 2024 | Unclear | Low | Low     | Unclear | 8:2 split stratified by BPPV type                              |
| Mun et al. 2024  | Unclear | Low | Unclear | High    | Patient-level split but clip-level analysis (unit-of-analysis) |

### Outcomes and effect measures

The primary outcome was diagnostic accuracy (proportion of correctly classified test units). Secondary outcomes were sensitivity, specificity, diagnostic odds ratio (DOR), positive and negative likelihood ratios, AUROC, the risk difference (RD) in accuracy between AI and clinicians reading the same cases, and the RD in clinician accuracy with versus without AI assistance. To avoid inflating precision when several readers interpreted the same images, clinician performance was expressed as an image-level equivalent (mean reader accuracy multiplied by the number of images).<sup>9,11</sup>

### Data synthesis

Proportions were logit-transformed and pooled using inverse-variance random-effects models because the Freeman–Tukey double-arcsine back-transformation can produce seriously misleading results. Between-study variance ( $\tau^2$ ) was estimated by restricted maximum likelihood (REML), which is preferred over the DerSimonian–Laird estimator, and 95% confidence intervals were computed with the modified Hartung–Knapp–Sidik–Jonkman (HKSJ) method, which maintains nominal coverage when studies are few or heterogeneous. Approximate 95% prediction intervals were derived from the t-distribution with  $k-2$  degrees of freedom. Studies reporting only a point estimate with a confidence interval contributed through the CI-derived standard error. AUROC values were pooled on their natural scale using CI-derived or Hanley–McNeil standard errors. RDs were pooled with the same random-effects approach.<sup>30–36</sup>

Heterogeneity was quantified using Cochran’s  $Q$ ,  $I^2$ ,  $\tau^2$  and prediction intervals. Prespecified subgroup analyses compared modality, population (paediatric vs adult/mixed), validation strategy (hold-out vs cross-validation; internal vs external/prospective/point-of-care), architecture family (single CNN, ensemble/hybrid deep learning, sequence/temporal models,

AutoML/classical machine learning), task complexity (binary, 3–4,  $\geq 5$  classes), reference standard (surgical/histopathological/objective vs expert annotation/clinical diagnosis), publication period, region and design, with between-subgroup heterogeneity tested by Q statistics. Mixed-effects meta-regression (REML with Knapp–Hartung adjustment) examined  $\log_{10}$  test-set size, publication year, number of classes and binary study-level covariates; the proportion of  $\tau^2$  explained ( $R^2$ ) was reported. Sensitivity and specificity were additionally summarised with a Moses–Littenberg summary receiver operating characteristic (SROC) curve.<sup>32,35</sup>

### **Small-study effects, influence and sensitivity analyses**

Small-study effects were explored using contour-enhanced funnel plots, Egger’s regression and Begg’s rank correlation, with Duval–Tweedie trim-and-fill as a sensitivity analysis; because conventional funnel plots can be misleading for proportions, a sample-size–based funnel plot was also inspected. Influence was evaluated with Galbraith radial plots, Baujat plots and leave-one-out analysis; temporal stability was explored with cumulative meta-analysis by publication year. Sensitivity analyses repeated the primary synthesis with DerSimonian–Laird and common-effect models and after excluding cross-validation estimates, approximated or reconstructed denominators, public/web-derived datasets, very large test sets, studies analysed at the image/clip level, and studies without surgical or objective reference standards.<sup>37,38</sup>

### **Certainty of evidence and software**

Certainty of evidence for each principal outcome was graded using the GRADE approach for diagnostic tests, considering risk of bias, inconsistency, indirectness, imprecision and publication bias. All analyses were implemented in Python 3.12 (NumPy 2.4, SciPy 1.17, Matplotlib 3.10) using custom, reproducible code for REML estimation, HKSJ inference, meta-regression and graphical diagnostics; a two-sided  $P < 0.05$  was considered statistically significant, with  $P < 0.10$  used for heterogeneity and small-study tests.<sup>33,37</sup>

---

## **RESULTS**

---

### **Characteristics of included studies**

The 64 included studies were published between 2019 and 2026 and originated predominantly from East Asia (China, Korea, Japan and Taiwan), followed by the United States, Europe, the Middle East, Southeast Asia, North Africa and Australia=. Thirty-seven studies addressed otoscopic or otoendoscopic diagnosis, 14 temporal bone CT, six audiometry or tympanometry and seven vestibular testing. Test-set sizes in meta-analysed accuracy studies ranged from 45 to 15,046 units (median 338; one study reported only a confidence interval).

Deep CNNs were the dominant architecture, increasingly combined in ensembles, hybrid segmentation–classification pipelines, three-dimensional CNNs for volumetric CT, and sequence models (LSTM, recurrent and one-dimensional CNNs) for audiometric and eye-movement signals; four studies evaluated general-purpose multimodal LLMs or automated machine learning. =<sup>39–45</sup>

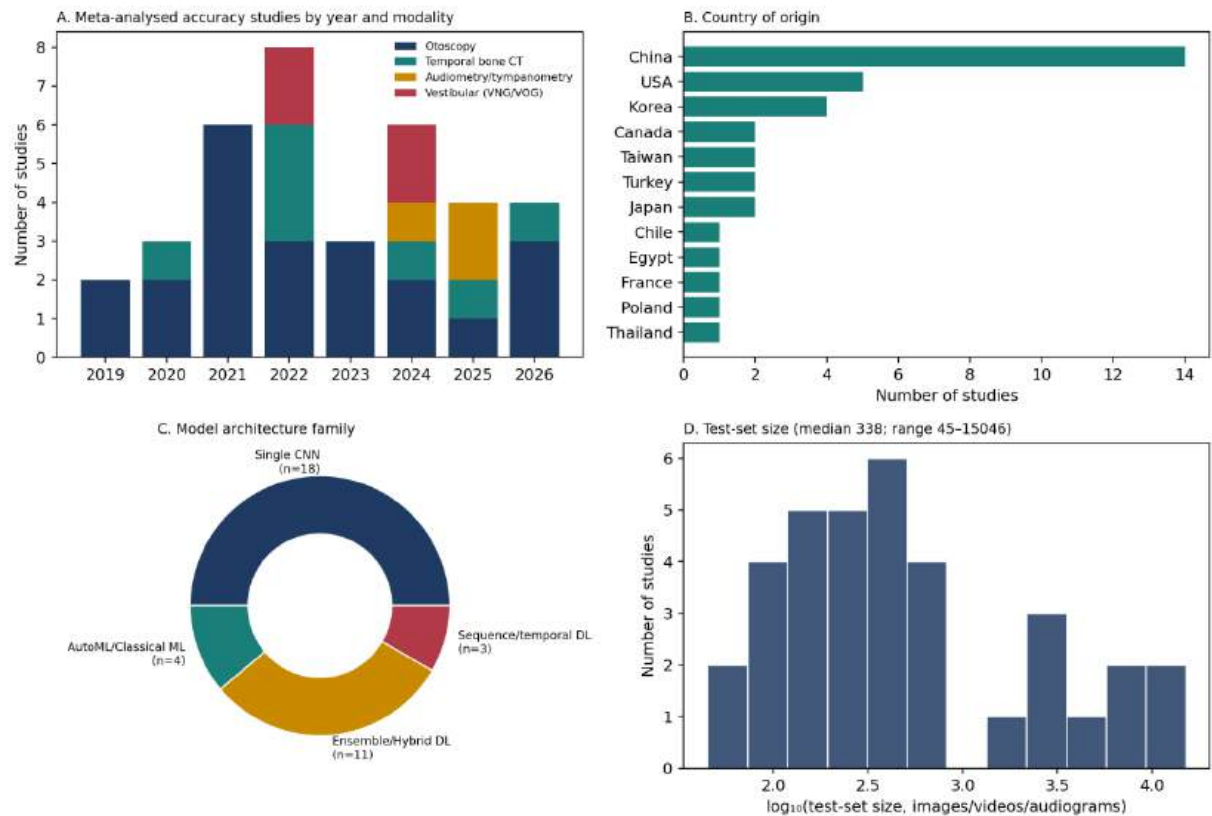


Figure 3. Characteristics of meta-analysed accuracy studies: (A) publications by year and modality; (B) country of origin; (C) model architecture family; (D) distribution of test-set sizes.

**Table 3. Characteristics of included studies — otoscopy and otoendoscopy (n = 37)**

| Study                          | Country | Design / setting                                      | Population; dataset      | Target condition(s) [classes]                                                  | AI model                          | Reference standard | Validation (test n)                  | Principal findings                                        |
|--------------------------------|---------|-------------------------------------------------------|--------------------------|--------------------------------------------------------------------------------|-----------------------------------|--------------------|--------------------------------------|-----------------------------------------------------------|
| Livingstone 2019 <sup>46</sup> | Canada  | Retrospective; single-centre image library            | Mixed ages; 734 images   | Normal TM, tympanostomy tube, cerumen impaction [3]                            | Custom 3-layer CNN                | Clinician labels   | Hold-out 20% (45)                    | Accuracy 84.4%                                            |
| Livingstone 2020 <sup>47</sup> | Canada  | Retrospective; web, open-repository and clinic images | Mixed; 1,366 images      | 14 otological diagnoses (multilabel) [14]                                      | Google Cloud AutoML Vision        | Consensus labels   | Hold-out (89) + physician comparison | Accuracy 88.7%; PPV 90.9%, recall 86.1%; physicians 58.9% |
| Cha 2019 <sup>39</sup>         | Korea   | Retrospective otoendoscopic database                  | Mixed; 10,544 images     | Normal, attic retraction, perforation, otitis externa ± myringitis, tumour [6] | Ensemble Inception-V3 + ResNet101 | Otologist labels   | 5-fold CV                            | Accuracy 93.67%                                           |
| Viscaino 2020 <sup>48</sup>    | Chile   | Retrospective                                         | 180 patients; 880 images | Normal, myringosclerosis, earwax plug, COM [4]                                 | SVM with filter-bank features     | Specialist labels  | Separate test (160)                  | Accuracy 93.9%; sensitivity 87.8%; specificity            |

|                            |       |                                                   |                                                          |                                                                              |                                          |                                     |                                                 |                                                                    |
|----------------------------|-------|---------------------------------------------------|----------------------------------------------------------|------------------------------------------------------------------------------|------------------------------------------|-------------------------------------|-------------------------------------------------|--------------------------------------------------------------------|
|                            |       |                                                   |                                                          |                                                                              |                                          |                                     |                                                 | 95.9%                                                              |
| Wu 2021 <sup>49</sup>      | China | Prospective (authors); institutional + smartphone | Paediatric; 12,203 otoendoscopic + 102 smartphone images | AOM, OME, normal [3]                                                         | Xception; MobileNet-V2                   | Specialist labels                   | Hold-out (1,500) + prospective smartphone (102) | Accuracy 97.45%; smartphone 90.66%                                 |
| Crowson 2021 <sup>50</sup> | USA   | Retrospective operative cohort                    | Children undergoing myringotomy; 338 images              | Middle ear effusion vs none [2]                                              | CNN                                      | Intraoperative myringotomy findings | 5-fold CV                                       | Accuracy 83.8%; AUROC 0.93; F1 0.80                                |
| Zeng X 2021 <sup>40</sup>  | China | Retrospective                                     | Mixed; 20,542 images                                     | Eight external/middle ear diseases incl. cholesteatoma, CSOM, otomycosis [8] | Ensemble DenseNet-BC169/BC1615           | Specialist labels                   | 80/20 split ( $\approx 4,108$ )                 | Accuracy 95.59%; real-time inference                               |
| Cai 2021 <sup>51</sup>     | China | Retrospective                                     | 2,022 participants; 6,066 images                         | Normal, OME, active and inactive CSOM [4]                                    | Two-stage attention-aware ResNet50 (CAM) | Specialist labels                   | 3-fold CV                                       | Accuracy 93.4%; performance comparable with an associate professor |
| Byun 2021 <sup>52</sup>    | Korea | Retrospective; reader study                       | Mixed; 2,272 images                                      | Normal, OME, COM,                                                            | ResNet18 + shuffle attention (Grad-      | Otologist labels                    | 71 representative                               | Accuracy 97.2%; resident                                           |

|                                |        |                                                  |                                                 |                                                          |                                                       |                                     |                                                                  |                                                                                                   |
|--------------------------------|--------|--------------------------------------------------|-------------------------------------------------|----------------------------------------------------------|-------------------------------------------------------|-------------------------------------|------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
|                                |        |                                                  |                                                 | cholesteatoma<br>[4]                                     | CAM)                                                  |                                     | cases; 10<br>residents                                           | accuracy 82.9%<br>→ 90.0% with<br>AI                                                              |
| Tsutsumi<br>2021 <sup>53</sup> | USA    | Retrospective;<br>public online<br>images        | 400 images                                      | Normal, AOM,<br>otitis externa,<br>CSOM, cerumen<br>[5]  | MobileNetV2 (web<br>deployment)                       | Image-source<br>labels              | 75:25 hold-<br>out (100)                                         | Binary AUROC<br>0.902; multiclass<br>accuracy 66%                                                 |
| Chen YC<br>2022 <sup>22</sup>  | Taiwan | Retrospective<br>2011–2019;<br>smartphone<br>app | Mixed; 2,820<br>images                          | Ten middle ear<br>conditions [10]                        | Nine CNNs;<br>ensembled<br>lightweight<br>MobileNetV2 | Three-specialist<br>consensus       | 8:2 split<br>(≈564); 50-<br>item reader<br>test                  | Accuracy 97.6%<br>(10-class),<br>98.0% (binary);<br>GPs 36%,<br>residents 80%,<br>specialists 90% |
| Crowson<br>2023 <sup>54</sup>  | USA    | Retrospective<br>cohort 2018–<br>2021            | Paediatric<br>operative; 639<br>training images | Normal, non-<br>purulent,<br>purulent effusion<br>[3]    | CNN; Google<br>AutoML<br>comparator                   | Intraoperative<br>findings          | Hold-out; 22-<br>image<br>clinician<br>survey (39<br>clinicians) | Accuracy 80.8%<br>(77.0–84.6);<br>survey: AI<br>95.5% vs<br>clinicians 65.0%                      |
| Mao<br>2022 <sup>55</sup>      | China  | Retrospective;<br>web<br>application             | ≈4,000 images<br>(2,182 train)                  | Otomycosis,<br>cerumen, otitis<br>externa, normal<br>[4] | EfficientNet-B6<br>(ensemble)                         | Fungal<br>culture/DNA<br>sequencing | Validation<br>(475)                                              | Accuracy<br>92.42%                                                                                |
| Zeng J                         | China  | Retrospective                                    | 1,239 participants                              | Conductive                                               | CNN vs logistic                                       | Pure-tone                           | 80/20 split                                                      | AUROC 0.74;                                                                                       |

|                              |           |                                                |                                                     |                                                        |                                                                |                               |                                    |                                                                                                |
|------------------------------|-----------|------------------------------------------------|-----------------------------------------------------|--------------------------------------------------------|----------------------------------------------------------------|-------------------------------|------------------------------------|------------------------------------------------------------------------------------------------|
| 2022 <sup>56</sup>           |           | multicentre<br>2015–2020                       | aged 4–89 y;<br>2,790 images                        | hearing loss in<br>OME (air–bone<br>gap > 10 dB) [2]   | regression                                                     | audiometry                    | (558)                              | accuracy 81%;<br>otologists 16–<br>39%                                                         |
| Zeng J<br>2023 <sup>57</sup> | China     | Retrospective,<br>three centres                | OME; 6,393<br>images                                | Attic retraction<br>pocket;<br>atelectasis [2<br>each] | CNN with CAM                                                   | Otologist panel               | 3-fold CV                          | ARP accuracy<br>89%, AUROC<br>0.89 (sens<br>0.93/spec 0.62);<br>atelectasis 79%,<br>AUROC 0.87 |
| Afify<br>2024 <sup>58</sup>  | Egypt     | Retrospective;<br>public<br>Chilean<br>dataset | 880 images                                          | Normal,<br>myringosclerosis,<br>earwax, COM [4]        | Bayesian-optimised<br>CNN                                      | Dataset labels                | Hold-out<br>(264)                  | Accuracy<br>98.10%;<br>sensitivity<br>98.11%;<br>specificity<br>99.36%                         |
| Shim<br>2023 <sup>59</sup>   | Korea     | Retrospective;<br>reader study                 | Paediatric; 1,150<br>+ 100<br>independent<br>images | OME vs normal<br>[2]                                   | ResNet50,<br>DenseNet201,<br>InceptionV3,<br>InceptionResNetV2 | Myringotomy-<br>confirmed OME | 5-fold CV +<br>independent<br>100  | DenseNet201<br>97.2%; seven<br>otolaryngologists<br>69–84%                                     |
| Suresh<br>2024 <sup>28</sup> | USA       | Prospective<br>point-of-care<br>comparison     | Children in<br>operative setting                    | Otitis media [2]                                       | Machine-learning<br>model                                      | Operative<br>findings         | Handheld<br>otoscopy,<br>same ears | AI 90.6% vs<br>paediatricians<br>59.4% (P = .01)                                               |
| Habib                        | Australia | Retrospective;                                 | 1,842 images                                        | Normal vs                                              | ResNet-50,                                                     | Cohort labels                 | Internal +                         | Mean AUROC                                                                                     |

|                           |        |                                                |                                                                      |                                         |                                                        |                       |                                     |                                                                                                      |
|---------------------------|--------|------------------------------------------------|----------------------------------------------------------------------|-----------------------------------------|--------------------------------------------------------|-----------------------|-------------------------------------|------------------------------------------------------------------------------------------------------|
| 2023 <sup>17</sup>        |        | three international public cohorts             | (Turkey, Chile, USA)                                                 | abnormal [2]                            | DenseNet-161, VGG16, ViT                               |                       | cross-cohort external               | 0.95 internal vs 0.76 external                                                                       |
| Dubois 2024 <sup>26</sup> | France | Retrospective ambulatory ENT 2013–2020         | Patients > 5 y; smartphone-attached otoscope; 41,664 training images | Eleven classes; normal vs abnormal [11] | Inception-v2 (iNside app)                              | Senior ENT specialist | Validation 3,962; held-out test 326 | Binary: 216/227 abnormal and 98/99 normal correctly classified; class-specific sensitivity 33.3–100% |
| Shaikh 2024 <sup>60</sup> | USA    | Prospective clinic videos 2018–2023, two sites | 635 children (mostly < 3 y); 1,151 smartphone videos                 | AOM vs no AOM [2]                       | Deep residual-recurrent network; decision-tree network | Validated otoscopists | Held-out test                       | Sensitivity 93.8%; specificity 93.5%; bulging aligned with AOM                                       |
| Noda 2024 <sup>44</sup>   | Japan  | Retrospective; reader comparison               | 305 images; 30 clinicians                                            | AOM, OME, COM, cholesteatoma [4]        | GPT-4 Vision with clinical data (zero-shot)            | Otologist diagnosis   | 190 test images                     | Accuracy 82.1%; paediatricians 70.6%; otolaryngologists > 95%                                        |
| Surapanen                 | USA    | Multicentre                                    | 219 children aged                                                    | Effusion vs no                          | Deep neural                                            | Myringotomy           | Not specified                       | Accuracy 92.1%;                                                                                      |

|                                    |          |                                          |                                         |                                             |                                                          |                                  |                            |                                                                          |
|------------------------------------|----------|------------------------------------------|-----------------------------------------|---------------------------------------------|----------------------------------------------------------|----------------------------------|----------------------------|--------------------------------------------------------------------------|
| i 2026 <sup>27</sup>               |          | (four institutions)                      | 6 mo–6 y; 537 images; consumer otoscope | effusion [2]                                | network                                                  | findings                         |                            | sensitivity 90.3%; specificity 93.8%                                     |
| Guo 2025 <sup>61</sup>             | China    | Retrospective                            | 819 images                              | Normal, AOM, OME, CSOM [4]                  | Five CNNs; VGG-19 best                                   | Specialist labels                | 60/20/20 split (≈164)      | Accuracy 94.51%; sensitivity 94.18%; specificity 98.10%                  |
| Yang 2026 <sup>62</sup>            | Taiwan   | Retrospective                            | 607 images                              | Normal, AOM, COM with perforation [3]       | U-Net segmentation + feature-based neural network        | Specialist labels                | Independent test (122)     | Accuracy 96.72%; mean Dice 71.68%                                        |
| Lv 2026 <sup>63</sup>              | China    | Retrospective 2016–2019; edge deployment | Paediatric; 19,522 images               | AOM, OME, normal [3]                        | MobileNetV3-Small (INT8 quantised)                       | Recorded diagnoses + findings    | Locked balanced test (600) | Accuracy 97.33% (INT8 97.17%); sensitivity 96–98%; specificity 98–99.75% |
| Prakairungthong 2026 <sup>64</sup> | Thailand | Prospective development and external     | 1,011 training images; 180 external     | Normal, OME, myringitis, perforation, other | ResNet classifiers + gradient-boosted model (mobile app) | Three-otolaryngologist consensus | Prospective external (180) | External accuracy 52.8%; κ 0.417;                                        |

|                                  |                |                                      |                                                          |                                        |                                             |                                     |                     |                                                             |
|----------------------------------|----------------|--------------------------------------|----------------------------------------------------------|----------------------------------------|---------------------------------------------|-------------------------------------|---------------------|-------------------------------------------------------------|
|                                  |                | validation                           |                                                          | [5]                                    |                                             |                                     |                     | perforation sensitivity 86.7%                               |
| Lu Hao<br>2026 <sup>65</sup>     | USA            | Multicentre; external paired cohort  | 104 otoscopy videos with tympanometry                    | Six middle ear categories [6]          | Video CNN + rule-guided tympanometry fusion | Clinical diagnosis                  | External (104)      | Video-only 63.5%; hybrid 81.7%                              |
| Ungkanont<br>2026 <sup>66</sup>  | Thailand       | Retrospective; web application       | Paediatric; 320 images                                   | OME vs normal [2]                      | InceptionV4                                 | Expert features + surgical findings | Internal validation | Accuracy 94.7%; AUROC 0.98 (0.93–1.00); $\kappa$ 0.627      |
| Sundgaard<br>2021* <sup>67</sup> | Denmark /Japan | Retrospective                        | 1,336 images                                             | AOM, OME, no effusion [3]              | Deep metric learning (triplet loss)         | Specialist labels                   | Cross-validation    | Precision comparable with expert otologists                 |
| Habib<br>2022* <sup>21</sup>     | Australia      | Retrospective tele-otology 2010–2018 | Aboriginal/Torres Strait Islander children; 6,527 images | AOM, COM, OME, wax/obstruction [multi] | Transfer-learning CNN                       | Otolaryngologists                   | Test (600)          | Accuracy 99.3% AOM, 96.3% COM, 77.8% OME; AUROC 0.963–0.997 |
| Choi<br>2022* <sup>68</sup>      | Korea          | Retrospective                        | Multi-label otoendoscopic database                       | OME, COM + secondary classes [multi]   | EfficientNet-B4 (customised)                | Specialist labels                   | Hold-out            | Primary-class Dice 95.19%; 2.6 ms per image                 |
| Park<br>2023* <sup>69</sup>      | Korea          | Retrospective                        | Otoscopy images                                          | Normal vs abnormal TM [2]              | Mask R-CNN substructure                     | Specialist labels                   | Hold-out            | Best AUROC 0.911 (malleus +                                 |

|                              |           |                                            |                                    |                                            |                                                                   |                    |                      |                                                                                                       |
|------------------------------|-----------|--------------------------------------------|------------------------------------|--------------------------------------------|-------------------------------------------------------------------|--------------------|----------------------|-------------------------------------------------------------------------------------------------------|
|                              |           |                                            |                                    |                                            | segmentation + classifier                                         |                    |                      | cone of light + umbo)                                                                                 |
| Chang 2024* <sup>70</sup>    | USA       | Prospective                                | 28 children; 123 smartphone images | Tympanostomy tube in place vs extruded [2] | Deep learning                                                     | Clinician otoscopy | Internal             | Accuracy 97.7%                                                                                        |
| Seo 2025* <sup>71</sup>      | Korea     | Retrospective                              | 2,964 images                       | Normal, AOM, OME, COM [4]                  | Four-stage pipeline (quality, segmentation, laterality, ConvNeXt) | Expert otologists  | Internal             | Accuracy 88.7%; quality classifier 98.8%                                                              |
| Lindmark 2026* <sup>72</sup> | Sweden    | Retrospective; five heterogeneous datasets | Multinational images               | Normal vs abnormal [2]                     | Multiple CNNs; factorial training design                          | Dataset labels     | External datasets    | External accuracy 83% (sens 71%, spec 91%) vs 53% baseline                                            |
| Xu 2026* <sup>18</sup>       | Australia | Retrospective; three public datasets       | Chile, Ohio and Türkiye datasets   | Normal vs abnormal [2]                     | CNN; counterfactual experiments                                   | Dataset labels     | Internal vs external | Dataset artefacts yield AUROC 0.89 internally and 0.87 externally from clinically irrelevant features |

*\*Qualitative synthesis only. AOM, acute otitis media; ARP, attic retraction pocket; CAM, class activation mapping; CNN, convolutional neural network; COM, chronic otitis media; CSOM, chronic suppurative otitis media; CV, cross-validation; OME, otitis media with effusion; PPV, positive predictive value; SVM, support vector machine; TM, tympanic membrane; ViT, vision transformer.*

**Table 4. Characteristics of included studies — temporal bone CT (n = 14)**

| Study                        | Country | Design / setting                     | Population; dataset              | Target condition(s) [classes]                    | AI model                                     | Reference standard     | Validation (test n)              | Principal findings                                                                        |
|------------------------------|---------|--------------------------------------|----------------------------------|--------------------------------------------------|----------------------------------------------|------------------------|----------------------------------|-------------------------------------------------------------------------------------------|
| Wang YM 2020 <sup>73</sup>   | China   | Retrospective; reader study          | 562 COM patients; 1,147 ears     | Normal, CSOM, cholesteatoma [3]                  | Region-proposal network + classification CNN | Ground-truth diagnosis | Hold-out 15% (172); six experts  | COM vs normal AUROC 0.92 (sens 83.3%/spec 91.4%); 3-class accuracy 76.7% vs experts 73.8% |
| Eroğlu 2022 <sup>74</sup>    | Turkey  | Retrospective case-control 2016–2021 | 300 patients (100 per group)     | COM with vs without cholesteatoma vs control [3] | Hybrid deep-learning model                   | Surgical findings      | 80/20 image split (3,093 images) | Accuracy 95.4%                                                                            |
| Takahashi 2022 <sup>75</sup> | Japan   | Retrospective surgical cohort        | 164 pars flaccida cholesteatomas | Mastoid extension vs none [2]                    | MobileNetV2 ensemble                         | Surgical findings      | 8-fold CV; 15 otolaryngologists  | Accuracy 81.1% (sens 85.0%/spec 77.3%) vs otolaryngologists 73.4%                         |
| Su 2022 <sup>76</sup>        | China   | Retrospective                        | 499 patients;                    | MEC, CSOM,                                       | Region                                       | Two-                   | Test 15% (271)                   | Accuracy 96.5%;                                                                           |

|                            |        |                                                    |                                                          |                                     |                                                       |                            |                                   |                                                                            |
|----------------------------|--------|----------------------------------------------------|----------------------------------------------------------|-------------------------------------|-------------------------------------------------------|----------------------------|-----------------------------------|----------------------------------------------------------------------------|
|                            |        | 2018–2020                                          | 998 ears                                                 | normal [3]                          | network + Siamese CNN + 3D-CNN                        | otolaryngologist consensus |                                   | mean AUROC 0.983                                                           |
| Chen BJ 2024 <sup>41</sup> | China  | Retrospective two-centre + prospective application | 1,661 + 108 patients (3,153 + 211 ears); 121 prospective | Pathological ear; cholesteatoma [2] | Explainable 3D-CNN (Grad-CAM)                         | Surgical findings          | CV, external and prospective      | AUROC 0.96 internal/0.93 external; accuracy 87.8%/84.3%; prospective 81.8% |
| Zhao 2025 <sup>77</sup>    | China  | Retrospective 2020–2021                            | 8,000 CT images                                          | Cholesteatoma vs normal [2]         | Weighted-voting ensemble (AlexNet, GoogLeNet, ResNet) | Histopathology             | 5-fold CV (200)                   | Accuracy 88.8%; sensitivity 89.8%; specificity 88.1%                       |
| Tang 2026 <sup>78</sup>    | China  | Retrospective; reader study                        | 1,200 patients; 2,011 images                             | Normal middle ear, OME, COM [3]     | Detection + classification CNN                        | Expert diagnosis           | Test (406); junior/senior readers | Accuracy 92.4%; readers 81.5% → 93.6% (junior), 87.9% → 95.6% (senior)     |
| Yağcı                      | Turkey | Retrospective                                      | 101 surgical                                             | Cholesteatoma vs                    | GPT-5 and                                             | Surgery +                  | Two sessions, 1                   | Accuracy 43.6–                                                             |

|                              |         |                            |                                  |                                                 |                                                          |                            |                           |                                                   |
|------------------------------|---------|----------------------------|----------------------------------|-------------------------------------------------|----------------------------------------------------------|----------------------------|---------------------------|---------------------------------------------------|
| 2026 <sup>45</sup>           |         | 2019–2024                  | patients                         | non-cholesteatomatous COM [2]                   | Gemini 2.5 Pro (zero-shot)                               | histopathology             | week apart                | 49.5% vs radiologist<br>96.0%; $\kappa$ 0.36–0.49 |
| Ouattassi 2025 <sup>79</sup> | Morocco | Retrospective case–control | 212 patients; external 125 cases | Cholesteatoma (partial middle-ear filling) [2]  | kNN, neural network, LR, SVM, RF                         | Surgical findings          | Internal + external (125) | External: neural network 4 errors (96.8%)         |
| Duan 2022 <sup>*80</sup>     | China   | Retrospective paediatric   | 119 patients                     | Cholesteatoma, LCH, middle ear inflammation [3] | VGG16_BN                                                 | Histopathology             | 70/30 split               | Accuracy 0.99 vs physicians 0.89–0.97             |
| Duan 2022b <sup>*81</sup>    | China   | Retrospective paediatric   | 61 patients                      | PCD-related OM vs simple OME [2]                | GoogLeNet and other CNNs                                 | Clinical/genetic diagnosis | Internal                  | Accuracy 0.99 (GoogLeNet)                         |
| Cao C 2023 <sup>*82</sup>    | China   | Retrospective              | Temporal bone CT patches         | CSOM, MEC, normal [3]                           | Graph isomorphism network (structure-constrained fusion) | Expert labels              | Internal                  | Accuracy 96.36%                                   |
| Wang Z 2022 <sup>*83</sup>   | China   | Retrospective              | 499 patients; 973 ears           | MEC, CSOM, normal [3]                           | MESIC (VGG-16)                                           | Otolaryngologist labels    | Internal                  | Mean AUROC 0.96; precision 90.1%; recall 85.4%    |

|                           |       |               |                    |                                       |                                                          |               |          |                 |
|---------------------------|-------|---------------|--------------------|---------------------------------------|----------------------------------------------------------|---------------|----------|-----------------|
| Qi<br>2024* <sup>84</sup> | China | Retrospective | 1,019 CT<br>images | CSOM,<br>cholesteatoma,<br>normal [3] | nnU-Net pre-<br>training +<br>transfer<br>classification | Expert labels | Internal | Accuracy 92.33% |
|---------------------------|-------|---------------|--------------------|---------------------------------------|----------------------------------------------------------|---------------|----------|-----------------|

\*Qualitative synthesis only. CSOM, chronic suppurative otitis media; LCH, Langerhans cell histiocytosis; LR, logistic regression; MEC, middle ear cholesteatoma; PCD, primary ciliary dyskinesia; RF, random forest.

**Table 5. Characteristics of included studies — audiometry and tympanometry (n = 6)**

| Study                              | Country  | Design /<br>setting                                            | Population;<br>dataset                                        | Target<br>condition(s)<br>[classes]                      | AI model                               | Reference<br>standard              | Validation<br>(test n)   | Principal findings                                                         |
|------------------------------------|----------|----------------------------------------------------------------|---------------------------------------------------------------|----------------------------------------------------------|----------------------------------------|------------------------------------|--------------------------|----------------------------------------------------------------------------|
| Kassjański<br>2024 <sup>42</sup>   | Poland   | Retrospective                                                  | 15,046 pure-<br>tone<br>audiograms                            | Normal,<br>conductive,<br>mixed,<br>sensorineural<br>[4] | Bidirectional<br>LSTM                  | Audiologist<br>classification      | Stratified<br>10-fold CV | Accuracy 99.33%                                                            |
| Jin 2023 <sup>85</sup>             | USA      | Data from<br>two cluster-<br>randomised<br>screening<br>trials | 1,635 rural<br>Alaskan<br>children;<br>4,810 tracing<br>pairs | Tympanogram<br>type A vs B/C<br>(pass/refer) [2]         | Hybrid deep<br>learning                | Audiologist<br>interpretation      | 10-fold CV               | Layperson data:<br>sensitivity 95.2%,<br>specificity 92.3%,<br>AUROC 0.968 |
| Achakulvisut<br>2025 <sup>86</sup> | Thailand | Retrospective                                                  | 1,959<br>audiogram<br>images (≥ 7 y)                          | Six hearing-<br>loss levels [6]                          | Object-<br>detection<br>digitisation + | Otolaryngologist<br>interpretation | 70/30 split<br>(352)     | Accuracy 96.43%                                                            |

|                            |            |                        |                                   |                                              |                                        |                            |                   |                                                 |
|----------------------------|------------|------------------------|-----------------------------------|----------------------------------------------|----------------------------------------|----------------------------|-------------------|-------------------------------------------------|
|                            |            |                        |                                   |                                              | random forest                          |                            |                   |                                                 |
| Disci 2025 <sup>87</sup>   | Turkey     | Retrospective          | 1,000 audiometric tests           | Hearing-loss type and degree [4]             | Artificial neural network (mobile app) | Audiologist classification | 800/200 split     | Type accuracy 94.5%; degree 95%                 |
| Crowson 2020 <sup>88</sup> | Canada/USA | Retrospective          | Adult audiogram reports           | Normal, conductive, mixed, sensorineural [4] | ResNet-101 (AutoAudio)                 | Audiologist labels         | Separate test set | Accuracy 97.5%; mixed losses most misclassified |
| Dou 2024 <sup>89</sup>     | China      | Retrospective (STROBE) | 6,259 patients; 12,518 audiograms | Degree, type, configuration                  | Deep learning framework                | Three audiologists         | Internal          | Accuracy 96.75–99.85% across tasks              |

\*Qualitative synthesis only. LSTM, long short-term memory.

**Table 6. Characteristics of included studies — vestibular testing (n = 7)**

| Study                    | Country | Design / setting                         | Population; dataset              | Target condition(s) [classes] | AI model                   | Reference standard  | Validation (test n) | Principal findings                                                     |
|--------------------------|---------|------------------------------------------|----------------------------------|-------------------------------|----------------------------|---------------------|---------------------|------------------------------------------------------------------------|
| Wagle 2022 <sup>90</sup> | USA     | Secondary analysis of AVERT trial videos | 435 video-oculography recordings | Nystagmus vs none [2]         | Ensemble ResNet/VGG (aEYE) | Single expert label | Cross-validation    | AUROC 0.86; sensitivity 88.4%; specificity 74.2%; folds accuracy 80.4% |

|                         |       |                                                    |                                           |                                             |                                                          |                                        |                                   |                                                                                    |
|-------------------------|-------|----------------------------------------------------|-------------------------------------------|---------------------------------------------|----------------------------------------------------------|----------------------------------------|-----------------------------------|------------------------------------------------------------------------------------|
| Anh 2022 <sup>91</sup>  | Japan | Retrospective consecutive neuro-otological work-up | 1,009 patients; 44 test features          | Peripheral vestibular vs non-peripheral [2] | RF, AdaBoost, GB, SVM, LR                                | Clinical diagnosis by neuro-otologists | 25% test (253)                    | Accuracy 76–79% (SVM 79%)                                                          |
| Wu P 2023 <sup>43</sup> | China | Retrospective development + separate test cohort   | 2,671 training; 703 test patients         | BPPV subtype classification [multi]         | 1D + deep-learning composite model                       | Clinical diagnosis                     | Separate test cohort              | Micro-AUROC 0.982; canal-specific AUROC 0.928–0.991                                |
| Lu H 2024 <sup>92</sup> | China | Retrospective                                      | 518 BPPV patients                         | Six BPPV categories [6]                     | Multimodal video + head-position cross-attention (BKTDN) | Clinical diagnosis                     | 8:2 split (≈104)                  | Accuracy 81.7% (unimodal 52.8%)                                                    |
| Mun 2024 <sup>93</sup>  | Korea | Retrospective                                      | VOG videos (360 test videos; 2,756 clips) | Nystagmus vs none [2]                       | 1D CNN on pupil trajectories                             | Expert labels                          | 8:1:1 patient-level split         | Accuracy 91.0%; sensitivity 94.1%; specificity 86.4%                               |
| Lim 2019 <sup>*94</sup> | Korea | Retrospective clinical data warehouse              | 3,467 patients; 91,778 videos             | Nystagmus direction; affected canal         | Deep neural network on grid images                       | Four otologists                        | Validation in 1,005 BPPV patients | Sensitivity/specificity 0.91/0.92 horizontal, 0.78/0.80 torsional; canal 0.81/0.97 |

|                             |       |               |                                 |                                                          |                                         |                  |          |                                        |                  |
|-----------------------------|-------|---------------|---------------------------------|----------------------------------------------------------|-----------------------------------------|------------------|----------|----------------------------------------|------------------|
| Kong<br>2022* <sup>95</sup> | China | Retrospective | Infrared<br>nystagmus<br>videos | BPPV<br>nystagmus<br>patterns;<br>torsional<br>nystagmus | ConvNeXt +<br>LSTM with<br>optical flow | Expert<br>labels | Internal | Accuracy<br>(patterns);<br>(torsional) | 94.91%<br>97.75% |
|-----------------------------|-------|---------------|---------------------------------|----------------------------------------------------------|-----------------------------------------|------------------|----------|----------------------------------------|------------------|

*\*Qualitative synthesis only. BPPV, benign paroxysmal positional vertigo; GB, gradient boosting; VOG, video-oculography.*

### **Pooled diagnostic accuracy across modalities**

Thirty-six internal-validation accuracy estimates were pooled. The overall pooled accuracy of task-specific AI was 92.5% (95% CI 90.0–94.5), with extreme heterogeneity ( $I^2 = 97.8\%$ ,  $\tau^2 = 0.858$ ;  $Q = 1593.2$ ,  $P < 0.001$ ) and a 95% prediction interval of 64.7–98.8%, indicating that the accuracy of a new AI system in a comparable setting could plausibly lie anywhere between moderate and near-perfect. Hypothesis H1 was therefore supported for the mean effect but not for the distribution of true effects.<sup>39,41,42,49</sup>

Accuracy differed significantly by modality (Q-between = 14.25,  $df = 3$ ,  $P = 0.003$ ). Otoscopic/otoendoscopic AI achieved 93.2% (95% CI 90.1–95.3) across 22 studies (PI 69.1–98.8%); temporal bone CT 90.2% (95% CI 81.6–95.0) across seven studies (PI 53.5–98.6%); audiometric and tympanometric interpretation 97.6% (95% CI 69.9–99.9) across three studies; and vestibular eye-movement analysis the lowest pooled accuracy, 84.0% (95% CI 70.7–91.9) across four studies (PI 35.8–98.0%). The lower vestibular estimate reflects the intrinsic difficulty of torsional and low-amplitude nystagmus detection, clip-level noise and the use of single-expert labels.<sup>42,90–93</sup>

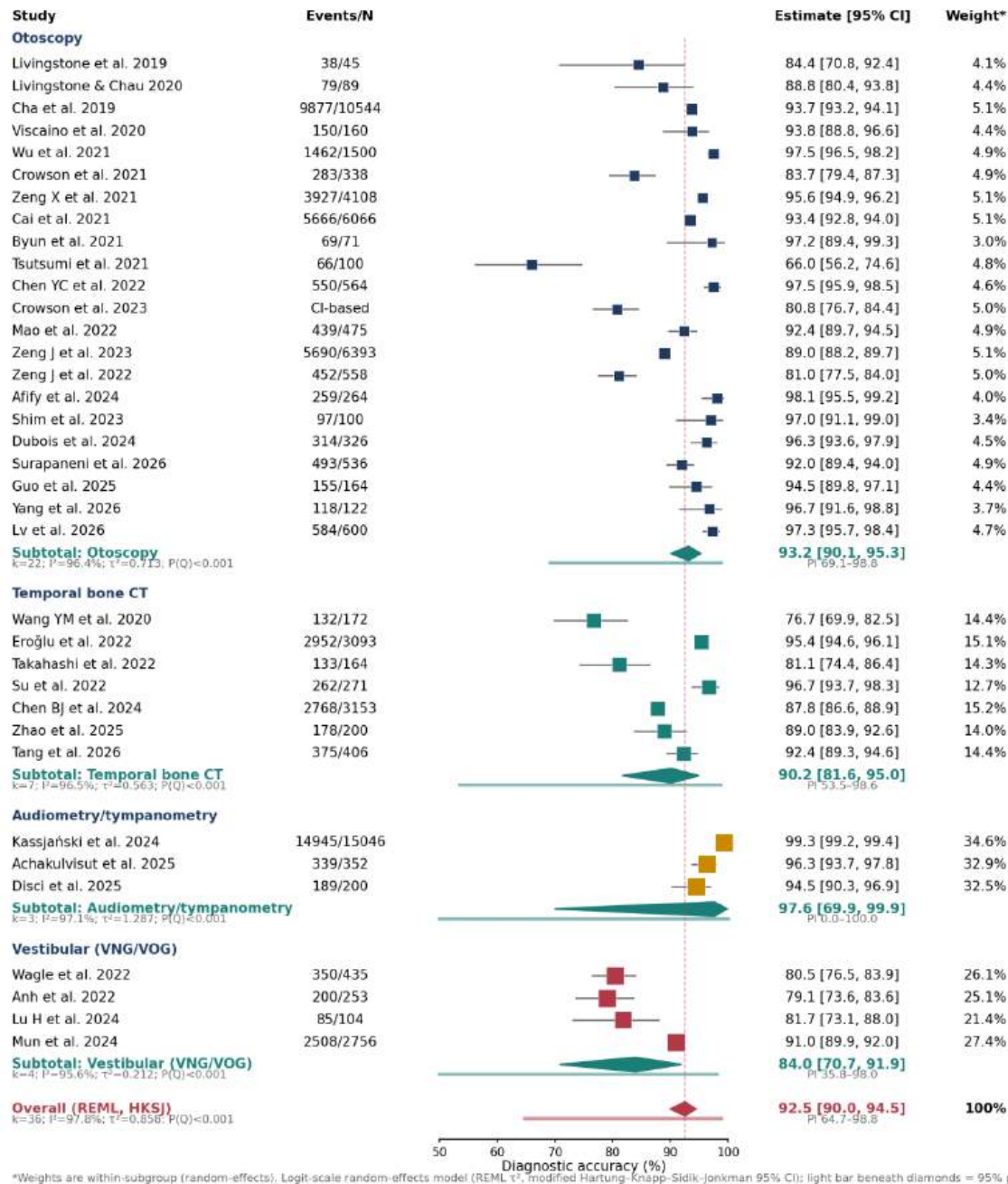


Figure 4. Forest plot of pooled diagnostic accuracy of task-specific AI by diagnostic otology modality (internal validation; k = 36). Squares are proportional to within-subgroup weight;

diamonds show pooled estimates with 95% CI; horizontal bars beneath the diamonds show 95% prediction intervals.

### **Subgroup analyses**

Pooled accuracy was similar in paediatric cohorts (93.5% (83.0–97.7;  $k = 6$ )) and adult or mixed-age cohorts (92.3% (89.4–94.5;  $k = 30$ );  $P = 0.700$ ), and in hold-out (92.9% (90.1–95.0;  $k = 27$ )) versus cross-validation estimates (91.2% (81.6–96.0;  $k = 9$ );  $P = 0.556$ ), although the cross-validation subgroup showed wider prediction intervals.<sup>27,49,50,59,63</sup>

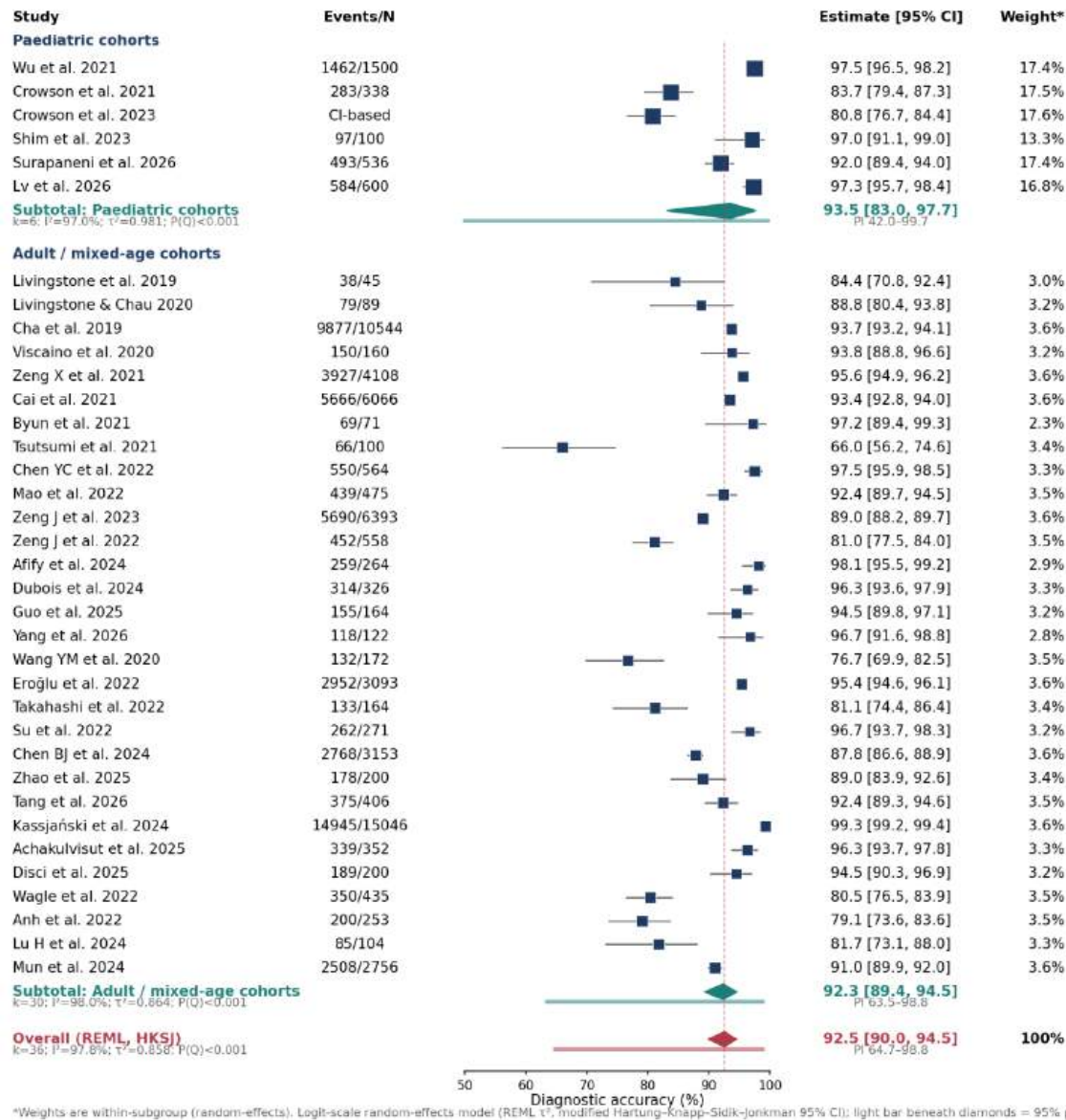


Figure 5. Forest plot of AI diagnostic accuracy stratified by population (paediatric vs adult/mixed-age cohorts).

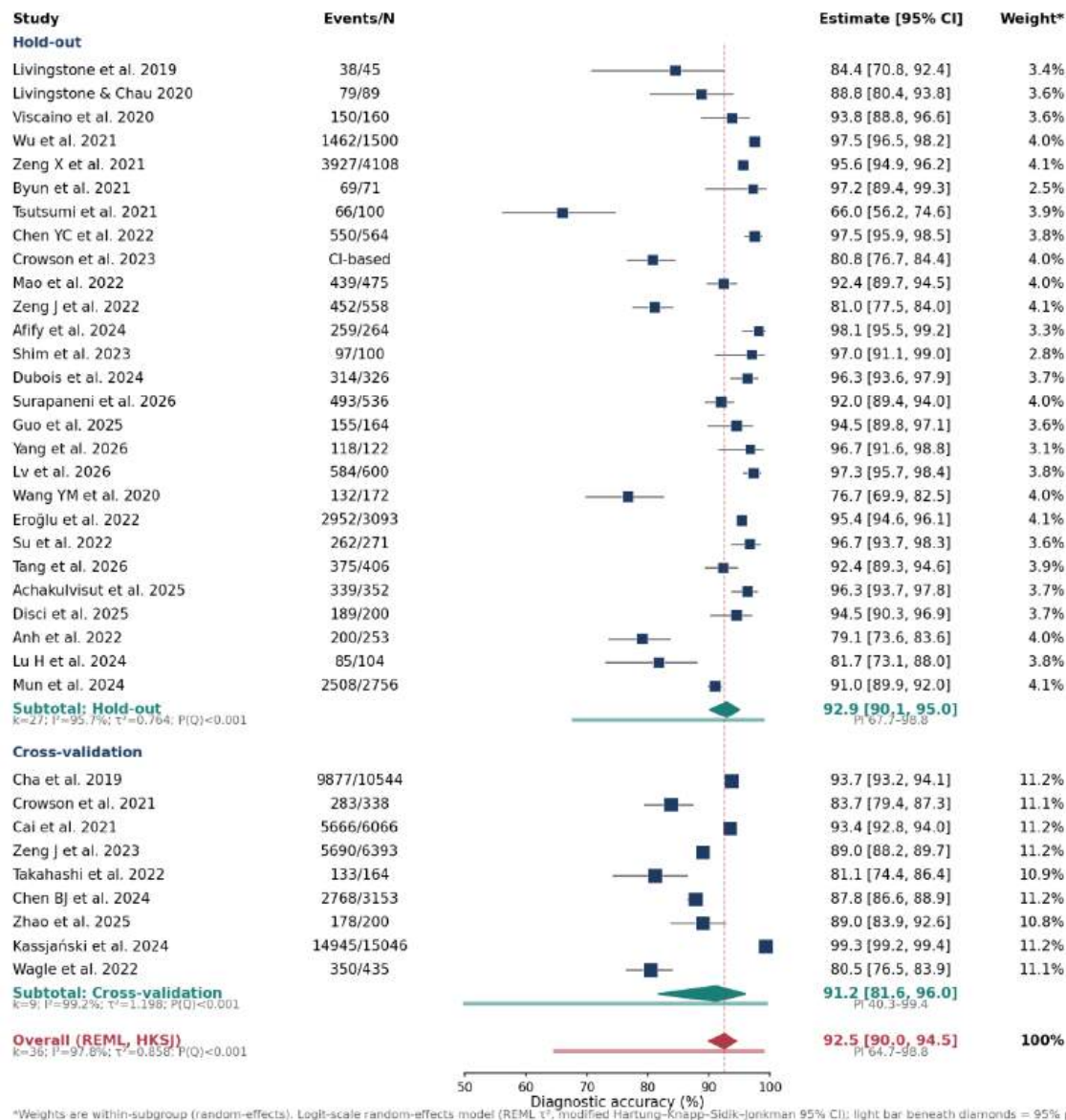


Figure 6. Forest plot of AI diagnostic accuracy stratified by internal validation strategy (independent hold-out test set vs k-fold cross-validation).

In contrast, performance declined materially outside development data. Seven external, prospective or point-of-care cohorts yielded a pooled accuracy of 81.2% (95% CI 59.1–92.8) with a prediction interval of 16.2–99.0%, significantly lower than internal validation (Q-between = 4.94, P = 0.026), supporting hypothesis H2. The decline ranged from minimal for an explainable three-

dimensional CT model tested at a second institute (84.3%) and a radiomic CT classifier (96.8%) to marked for smartphone applications and video models in new populations (52.8–63.5%).<sup>17,28,41,49,64,65,79</sup>

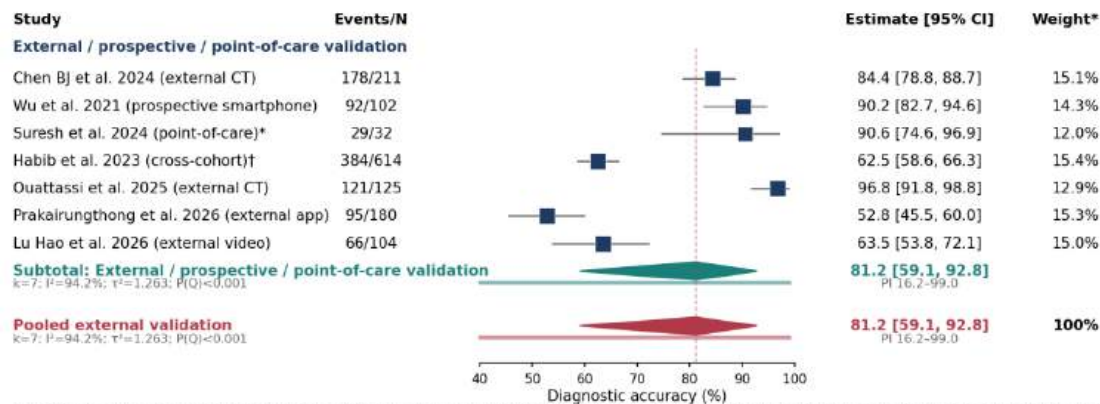


Figure 7. Forest plot of AI diagnostic accuracy in external, prospective or point-of-care validation cohorts.

Architecture family did not explain heterogeneity (single CNN 92.5% (88.5–95.2;  $k = 18$ ); ensemble/hybrid deep learning 92.4% (87.2–95.6;  $k = 11$ ); sequence/temporal models 95.0% (16.7–99.9;  $k = 3$ ); AutoML/classical machine learning 90.0% (73.8–96.7;  $k = 4$ );  $P = 0.853$ ). Task complexity was associated with accuracy ( $P = 0.010$ ), but non-monotonically: binary tasks yielded the lowest pooled accuracy (88.1% (83.1–91.7;  $k = 12$ )), reflecting that many binary tasks targeted subtle, clinically demanding distinctions (effusion vs no effusion, conductive hearing loss, mastoid extension, nystagmus detection), whereas three- to four-class tasks reached 94.9% (91.8–96.8;  $k = 17$ ) and tasks with five or more classes 91.8% (80.8–96.8;  $k = 7$ ).<sup>50,56,75,90</sup>

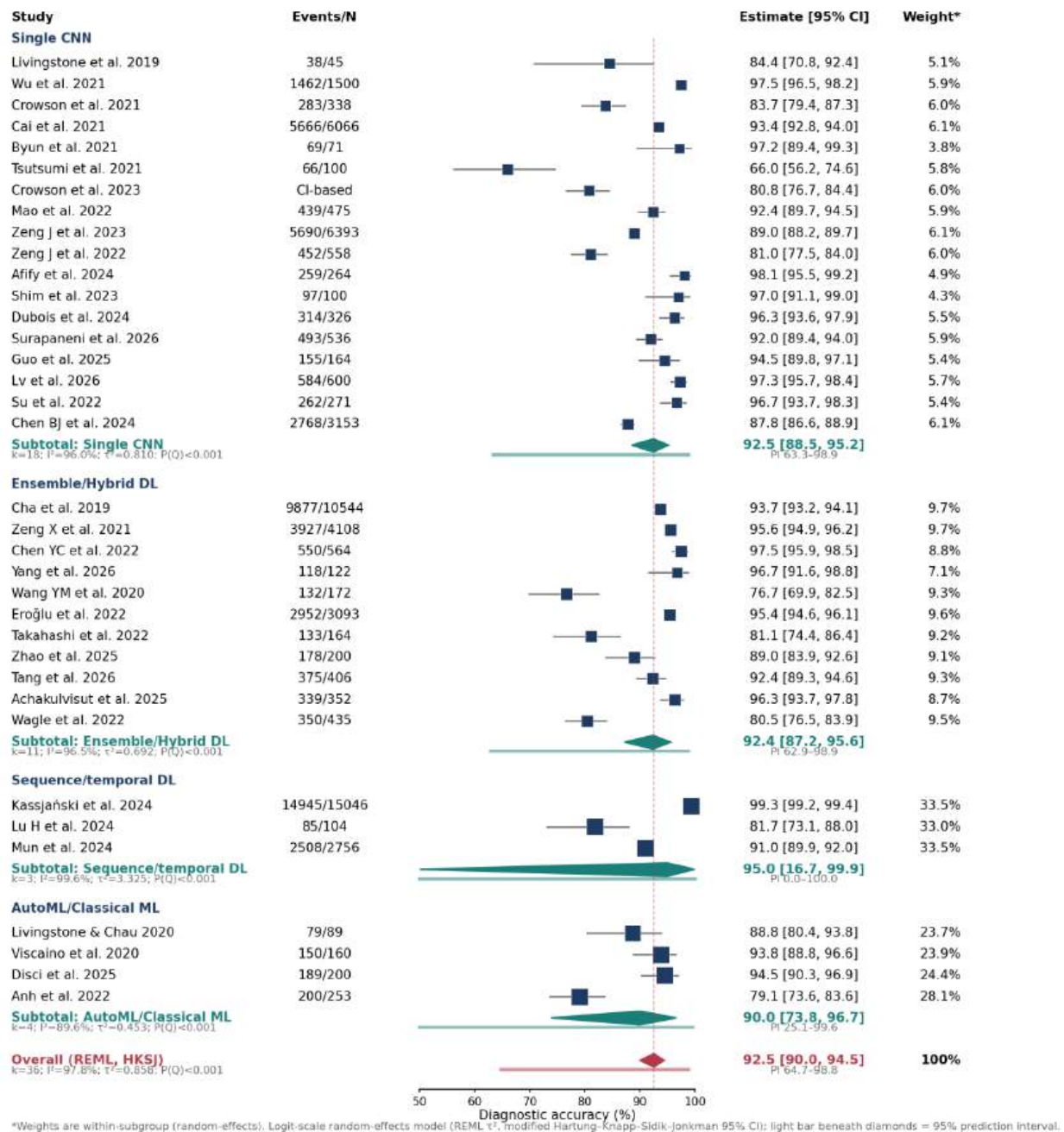


Figure 8. Forest plot of AI diagnostic accuracy stratified by model architecture family.

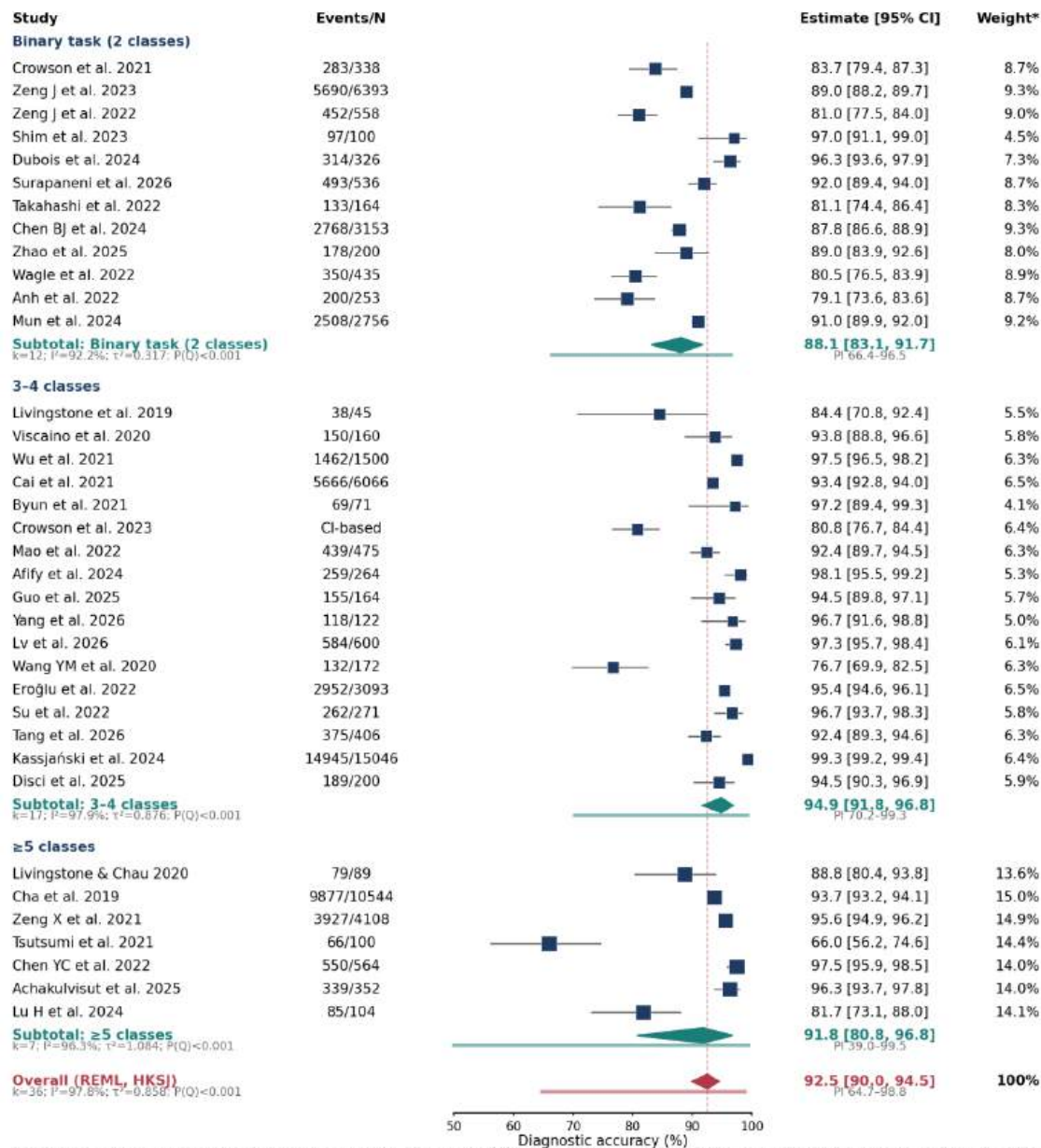


Figure 9. Forest plot of AI diagnostic accuracy stratified by task complexity (number of diagnostic classes).

The reference standard was the most informative design-related moderator. Studies anchored to surgical, myringotomy, histopathological, mycological or audiometric references reported a pooled accuracy of 89.0% (83.4–92.8;  $k = 10$ ), significantly lower than studies relying on

expert annotation or clinical diagnosis (93.6% (90.6–95.7;  $k = 26$ );  $Q$ -between = 4.18,  $P = 0.041$ ), consistent with hypothesis H4. Accuracy increased modestly in the most recent publication period (2024–2026: 94.8% (91.4–96.9;  $k = 14$ ) vs 2019–2021: 90.9% (84.0–95.0;  $k = 11$ );  $P = 0.139$ ), and did not differ between East Asian and other regions ( $P = 0.964$ ). Only two meta-analysed internal-validation studies used prospective data collection, precluding a meaningful design comparison ( $P = 0.374$ ).<sup>27,55,59,75,77</sup>

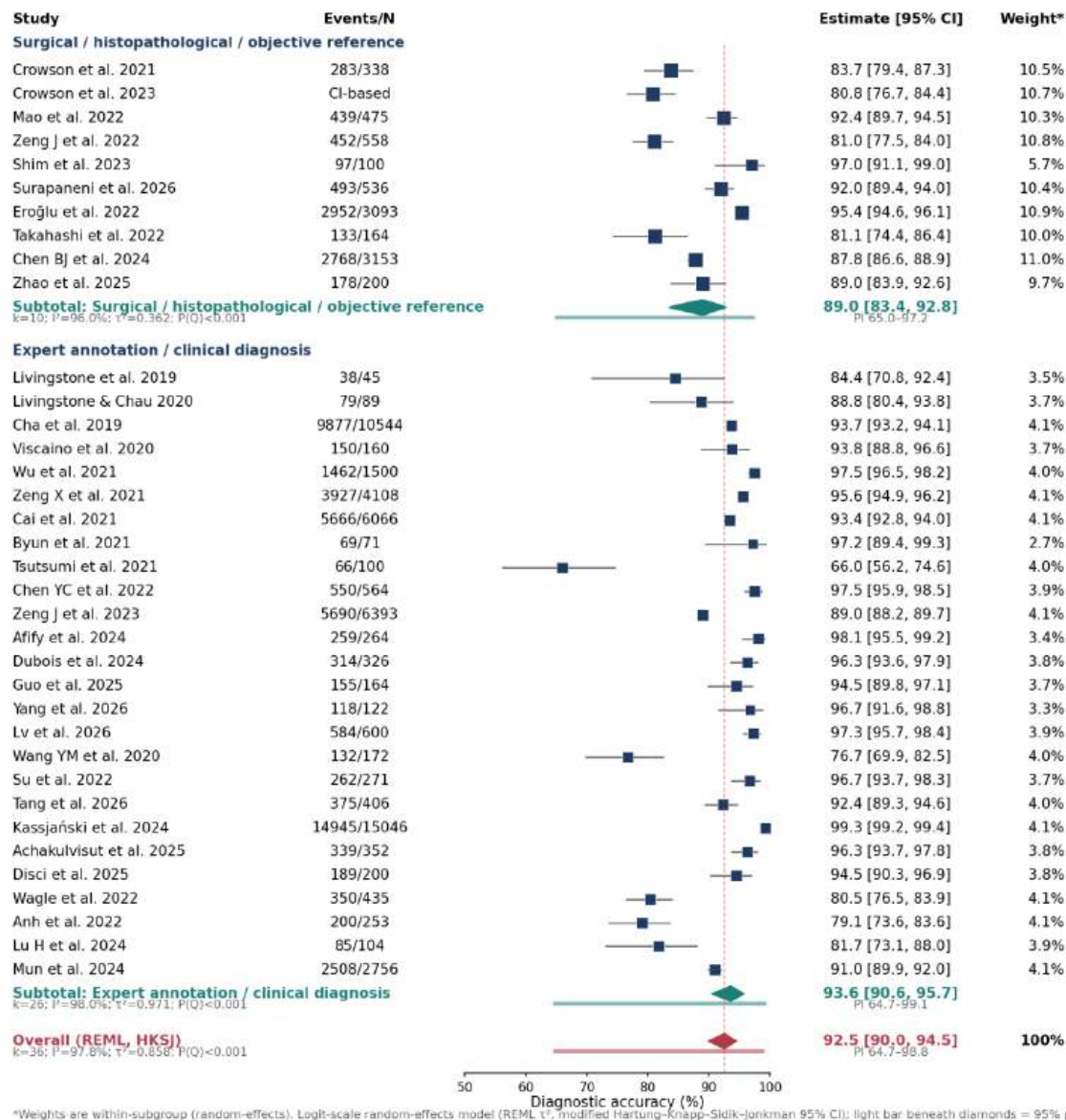


Figure 10. Forest plot of AI diagnostic accuracy stratified by reference standard (surgical/histopathological/objective vs expert annotation or clinical diagnosis).

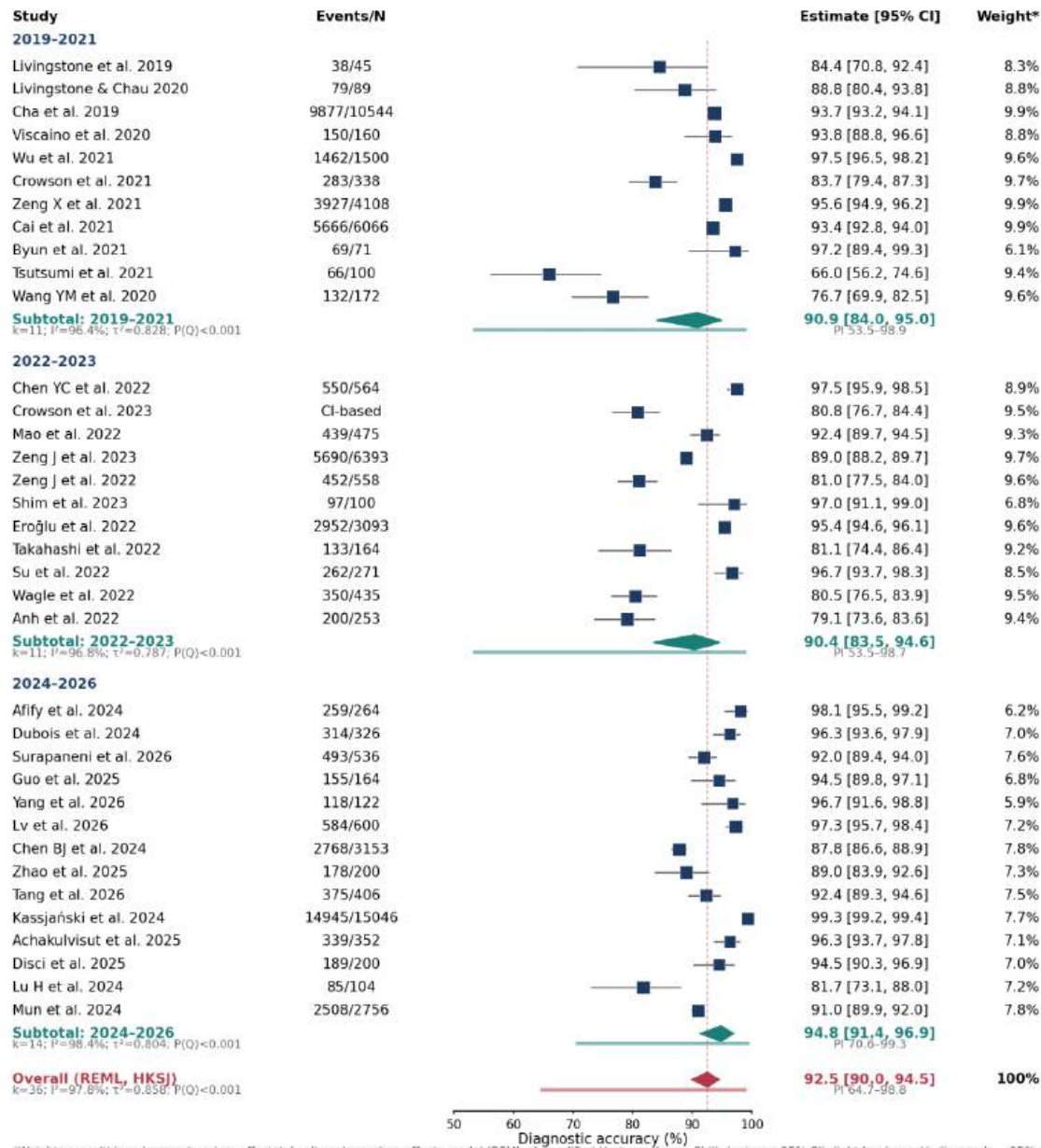


Figure 11. Forest plot of AI diagnostic accuracy stratified by publication period.

Table 7. Subgroup analyses of pooled AI diagnostic accuracy (logit random-effects, REML, modified HKSJ)

| Subgroup | k | Accuracy, % | 95% PI, | I <sup>2</sup> , % | Test for |
|----------|---|-------------|---------|--------------------|----------|
|----------|---|-------------|---------|--------------------|----------|

|                                                |    | (95% CI)         | %         |      | subgroup differences                  |
|------------------------------------------------|----|------------------|-----------|------|---------------------------------------|
| Modality: Otoscopy                             | 22 | 93.2 (90.1–95.3) | 69.1–98.8 | 96.4 | Q=14.25; df=3; P = 0.003              |
| Modality: Temporal bone CT                     | 7  | 90.2 (81.6–95.0) | 53.5–98.6 | 96.5 |                                       |
| Modality: Audiometry/tympanometry              | 3  | 97.6 (69.9–99.9) | 0.0–100.0 | 97.1 |                                       |
| Modality: Vestibular (VNG/VOG)                 | 4  | 84.0 (70.7–91.9) | 35.8–98.0 | 95.6 |                                       |
| Population: Paediatric cohorts                 | 6  | 93.5 (83.0–97.7) | 42.0–99.7 | 97.0 | Q=0.15; df=1; P = 0.700               |
| Population: Adult / mixed-age cohorts          | 30 | 92.3 (89.4–94.5) | 63.5–98.8 | 98.0 |                                       |
| Validation: Hold-out                           | 27 | 92.9 (90.1–95.0) | 67.7–98.8 | 95.7 | Q=0.35; df=1; P = 0.556               |
| Validation: Cross-validation                   | 9  | 91.2 (81.6–96.0) | 40.3–99.4 | 99.2 |                                       |
| Validation: external/prospective/point-of-care | 7  | 81.2 (59.1–92.8) | 16.2–99.0 | 94.2 | Q=4.94; df=1; P = 0.026 (vs internal) |
| Architecture: Single CNN                       | 18 | 92.5 (88.5–95.2) | 63.3–98.9 | 96.0 | Q=0.78; df=3; P = 0.853               |
| Architecture: Ensemble/Hybrid DL               | 11 | 92.4 (87.2–95.6) | 62.9–98.9 | 96.5 |                                       |
| Architecture: Sequence/temporal DL             | 3  | 95.0 (16.7–99.9) | 0.0–100.0 | 99.6 |                                       |

|                                                                     |    |                      |           |      |                            |
|---------------------------------------------------------------------|----|----------------------|-----------|------|----------------------------|
| Architecture:<br>AutoML/Classical ML                                | 4  | 90.0 (73.8–<br>96.7) | 25.1–99.6 | 89.6 |                            |
| Task: Binary task (2 classes)                                       | 12 | 88.1 (83.1–<br>91.7) | 66.4–96.5 | 92.2 | Q=9.19; df=2; P<br>= 0.010 |
| Task: 3–4 classes                                                   | 17 | 94.9 (91.8–<br>96.8) | 70.2–99.3 | 97.9 |                            |
| Task: $\geq 5$ classes                                              | 7  | 91.8 (80.8–<br>96.8) | 39.0–99.5 | 96.3 |                            |
| Reference: Surgical /<br>histopathological / objective<br>reference | 10 | 89.0 (83.4–<br>92.8) | 65.0–97.2 | 96.0 | Q=4.18; df=1; P<br>= 0.041 |
| Reference: Expert annotation /<br>clinical diagnosis                | 26 | 93.6 (90.6–<br>95.7) | 64.7–99.1 | 98.0 |                            |
| Period: 2019–2021                                                   | 11 | 90.9 (84.0–<br>95.0) | 53.5–98.9 | 96.4 | Q=3.95; df=2; P<br>= 0.139 |
| Period: 2022–2023                                                   | 11 | 90.4 (83.5–<br>94.6) | 53.5–98.7 | 96.8 |                            |
| Period: 2024–2026                                                   | 14 | 94.8 (91.4–<br>96.9) | 70.6–99.3 | 98.4 |                            |
| Region: East Asia                                                   | 22 | 92.5 (89.5–<br>94.7) | 69.5–98.5 | 96.5 | Q=0.00; df=1; P<br>= 0.964 |
| Region: Other regions                                               | 14 | 92.6 (86.4–<br>96.1) | 48.8–99.4 | 98.6 |                            |

### Meta-regression

In univariable mixed-effects meta-regression, larger test sets were associated with higher accuracy ( $\beta = 0.486$  logit units per  $\log_{10}$  increase;  $P = 0.047$ ;  $R^2 = 11.4\%$ ), a pattern compatible with greater development data, more homogeneous single-institution archives and, potentially, near-

duplicate images. Publication year showed a non-significant positive trend ( $\beta = 0.135$ ;  $P = 0.094$ ), and the number of classes, paediatric population, cross-validation, otoscopic modality and surgical/objective reference standard did not individually reach significance; the latter explained 5.1% of  $\tau^2$ . Residual heterogeneity remained substantial in every model, indicating that most between-study variation is attributable to unmeasured dataset- and label-level factors.<sup>17,18,26</sup>

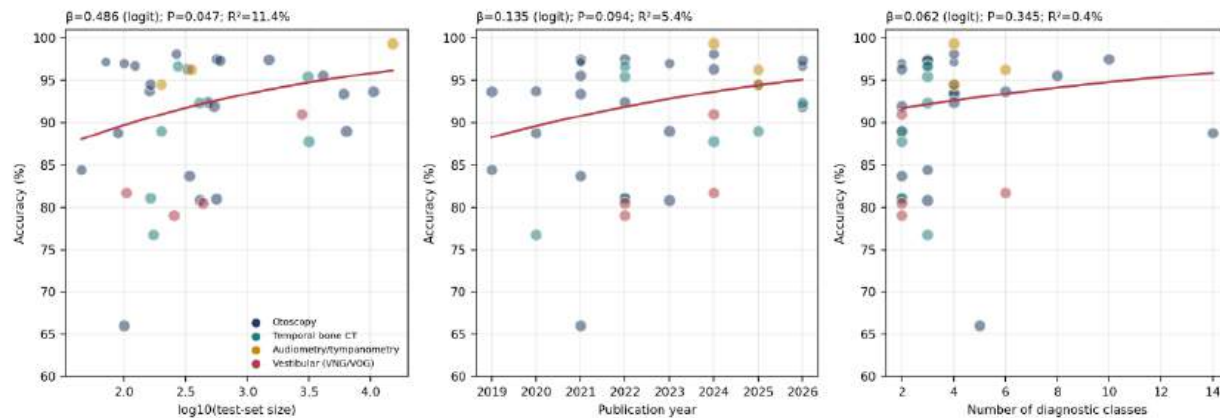


Figure 12. Bubble plots of univariable meta-regression of accuracy on (left)  $\log_{10}$  test-set size, (centre) publication year and (right) number of diagnostic classes. Bubble size is proportional to random-effects weight; colours denote modality; the red line shows the fitted logit-linear relationship.

**Table 8. Univariable mixed-effects meta-regression of logit accuracy (k = 36; REML with Knapp–Hartung adjustment)**

| Covariate                            | $\beta$<br>(logit) | SE    | P value | Residual<br>$\tau^2$ | $R^2$ , % |
|--------------------------------------|--------------------|-------|---------|----------------------|-----------|
| $\log_{10}(\text{test-set size})$    | 0.486              | 0.235 | 0.047   | 0.760                | 11.4      |
| Publication year                     | 0.135              | 0.078 | 0.094   | 0.812                | 5.4       |
| Number of diagnostic classes         | 0.062              | 0.065 | 0.345   | 0.855                | 0.4       |
| Paediatric cohort (1=yes)            | 0.171              | 0.436 | 0.697   | 0.881                | 0.0       |
| Cross-validation (1=yes)             | -0.243             | 0.367 | 0.513   | 0.877                | 0.0       |
| Surgical/objective reference (1=yes) | -0.564             | 0.347 | 0.113   | 0.814                | 5.1       |

|                                              |       |       |       |       |     |
|----------------------------------------------|-------|-------|-------|-------|-----|
| Otoscopy vs other modalities<br>(1=otoscopy) | 0.245 | 0.330 | 0.462 | 0.873 | 0.0 |
|----------------------------------------------|-------|-------|-------|-------|-----|

### Sensitivity, specificity, diagnostic odds ratio and AUROC

Seven studies provided disease-oriented sensitivity and specificity. Pooled sensitivity was 92.6% (95% CI 89.4–94.8) with moderate heterogeneity ( $I^2 = 67.8\%$ ,  $\tau^2 = 0.108$ ; PI 82.9–97.0%), whereas pooled specificity was 89.5% (95% CI 74.3–96.1) with extreme heterogeneity ( $I^2 = 99.6\%$ ,  $\tau^2 = 1.070$ ; PI 32.0–99.3%), driven by the low specificity of attic retraction pocket detection (62.0%) and mastoid-extension prediction (77.4%). The pooled diagnostic odds ratio was 104.7 (95% CI 26.8–408.3), corresponding to a positive likelihood ratio of 8.8 and a negative likelihood ratio of 0.08. The Moses–Littenberg SROC curve yielded an area of 0.908.<sup>26,27,57,60,75,77,85</sup>

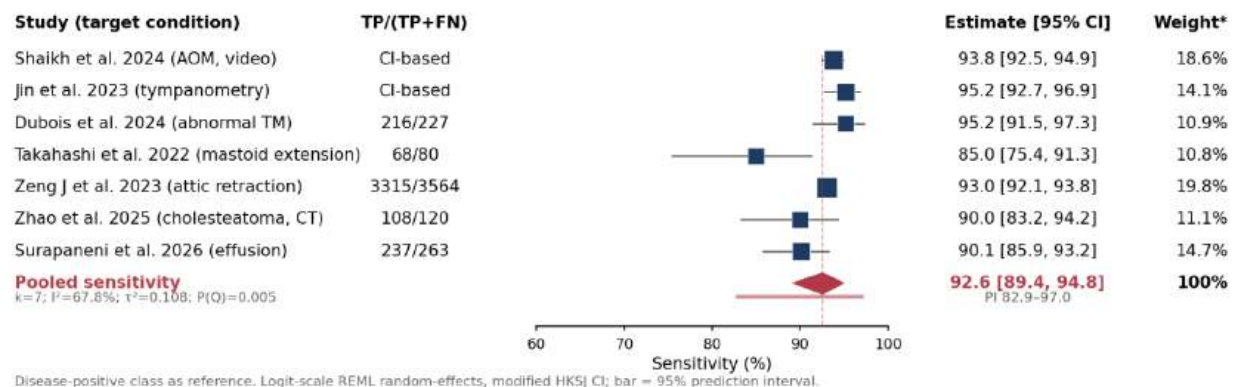


Figure 13. Forest plot of pooled sensitivity of AI for otological target conditions (disease-positive class as reference).

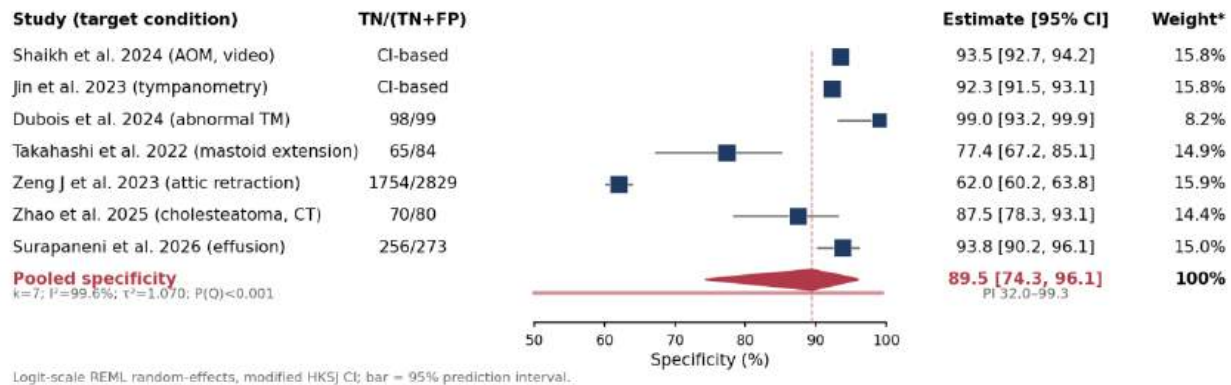


Figure 14. Forest plot of pooled specificity of AI for otological target conditions.

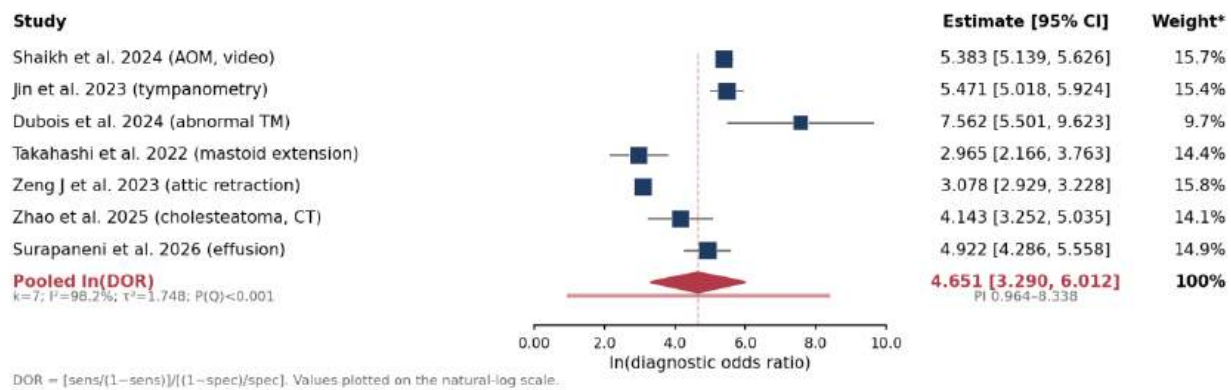


Figure 15. Forest plot of the natural logarithm of the diagnostic odds ratio.

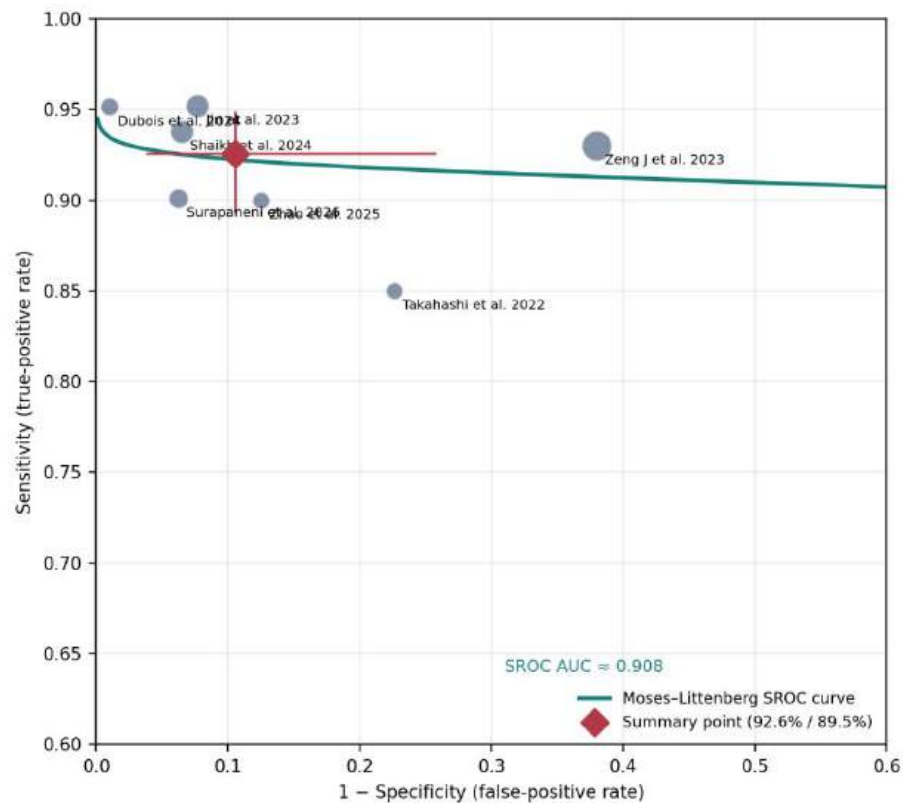


Figure 16. Summary receiver operating characteristic (SROC) plot. Circles represent individual studies (size proportional to sample size); the red diamond indicates the summary sensitivity and false-positive rate with 95% CI; the teal line is the Moses–Littenberg SROC curve.

Six studies reported AUROC with a computable standard error. The pooled AUROC was 0.957 (95% CI 0.914–1.000;  $I^2 = 99.0\%$ ), ranging from 0.89 for attic retraction pocket detection to  $\geq 0.98$  for BPPV subtype classification, paediatric OME and normal-versus-abnormal smartphone otoscopy.<sup>26,43,50,57,66,85</sup>

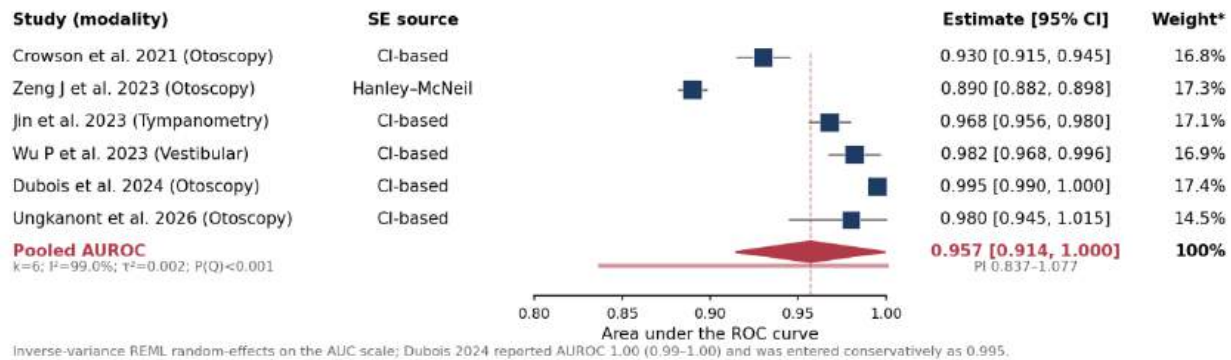


Figure 17. Forest plot of pooled area under the receiver operating characteristic curve (AUROC).

Applied to 1,000 patients, the pooled sensitivity and specificity imply that at a prevalence of 40%—typical of referred otological populations—AI would correctly identify approximately 370 of 400 diseased ears, miss 30, and falsely flag 63 of 600 non-diseased ears. The wide range of false positives (from 23 to 154 per 1,000) reflects the imprecision of the pooled specificity and is clinically relevant for antibiotic stewardship and referral capacity.<sup>60,85</sup>

**Table 9. Expected classification per 1,000 tested ears based on pooled sensitivity (92.6%) and specificity (89.5%)**

| Prevalence | True positives | False negatives<br>(95% CI) | True negatives | False positives<br>(95% CI) |
|------------|----------------|-----------------------------|----------------|-----------------------------|
| 20%        | 185            | 15 (10–21)                  | 716            | 84 (31–206)                 |
| 40%        | 370            | 30 (21–43)                  | 537            | 63 (23–154)                 |
| 60%        | 555            | 45 (31–64)                  | 358            | 42 (15–103)                 |

### AI compared with clinicians and AI-assisted interpretation

Nine studies compared AI with clinicians on identical cases. Task-specific AI was more accurate than clinicians by 18.2 percentage points (95% CI 7.3 to 29.0; k = 7; I<sup>2</sup> = 73.1%), supporting hypothesis H3. The advantage was largest against non-specialists—general physicians, paediatricians and mixed primary care clinicians (RD +30 to +32 points)—and smallest against otologists reading temporal bone CT (RD +3 to +8 points). By contrast, general-purpose multimodal LLMs were on average less accurate than clinicians (RD -25.0 points, 95% CI -72.7 to

22.7;  $k = 4$ ); GPT-4 Vision outperformed paediatricians but not otolaryngologists on otoscopic images, while GPT-5 and Gemini 2.5 Pro performed at near-chance level on temporal bone CT. The difference between task-specific AI and LLMs was significant (Q-between = 7.63,  $P = 0.006$ ).<sup>28,44,45,47,52,54,59,73,75</sup>

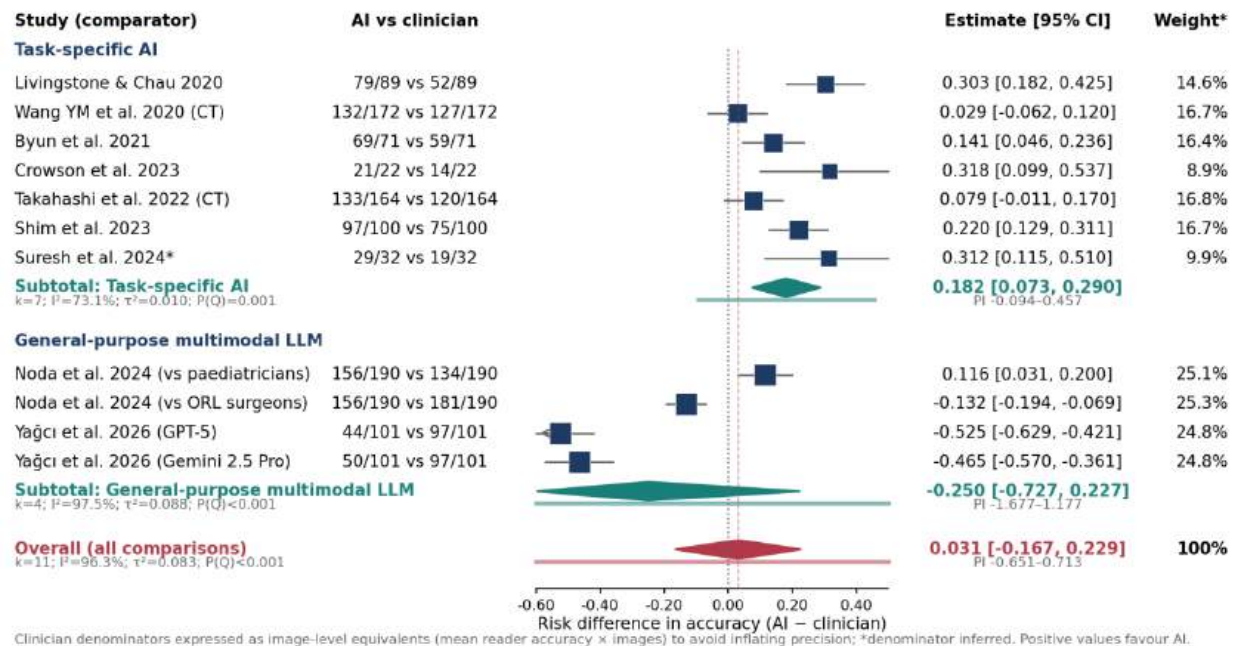


Figure 18. Forest plot of the risk difference in diagnostic accuracy between AI and clinicians reading identical cases, stratified by task-specific AI versus general-purpose multimodal large language models.

Three reader strata from two studies quantified the incremental value of AI assistance. Clinician accuracy increased by 9.4 percentage points with AI support (95% CI 1.4 to 17.4;  $I^2 = 16.4\%$ ), with the largest gain among junior clinicians interpreting temporal bone CT (81.5% to 93.6%) and among otolaryngology residents interpreting tympanic membrane images (82.9% to 90.0%). Residents tended to retain a correct initial diagnosis when the network was wrong, suggesting appropriate rather than automation-biased reliance in that setting.<sup>52,78</sup>

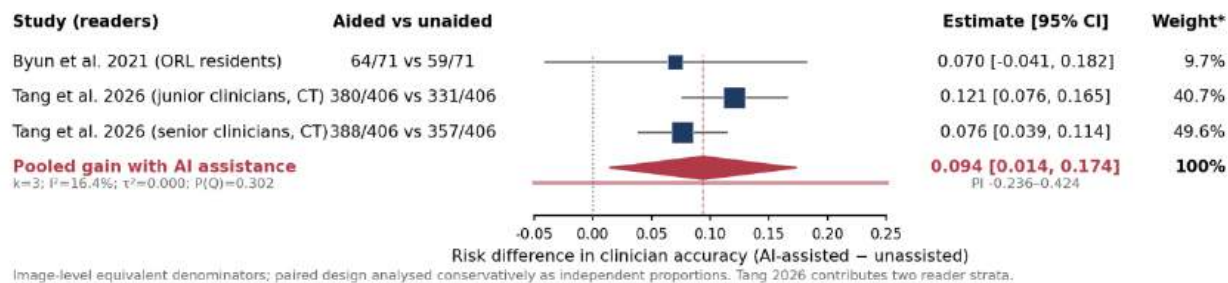


Figure 19. Forest plot of the change in clinician accuracy with versus without AI assistance (risk difference).

Pooled accuracy of zero-shot multimodal LLMs was 60.3% (95% CI 12.0–94.4) ( $k = 3$ ; PI 0.0–100.0%), substantially below task-specific models. Between-session reproducibility of LLM outputs on identical CT images was only fair to moderate ( $\kappa$  0.36–0.49), and inter-model agreement was slight, highlighting stochasticity as an additional barrier for general-purpose models.<sup>44,45</sup>

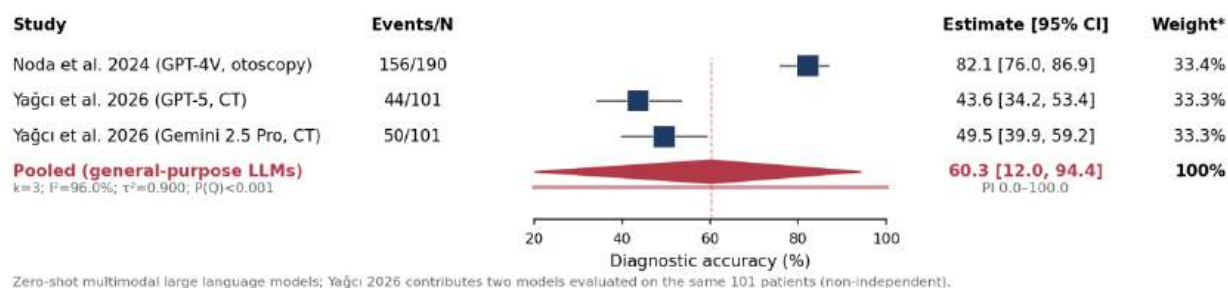


Figure 20. Forest plot of diagnostic accuracy of general-purpose multimodal large language models (zero-shot).

Figure 22 contrasts AI with clinicians of different experience levels. Across studies, a consistent experience gradient emerged: general practitioners and paediatricians performed worst (36–71%), residents intermediate (70–83%), and otology specialists best (82.5–95%), with task-specific AI equalling or exceeding specialists on curated test sets and narrowing the gap between junior and senior clinicians when used as an assistant.<sup>22,44,52,59,78</sup>

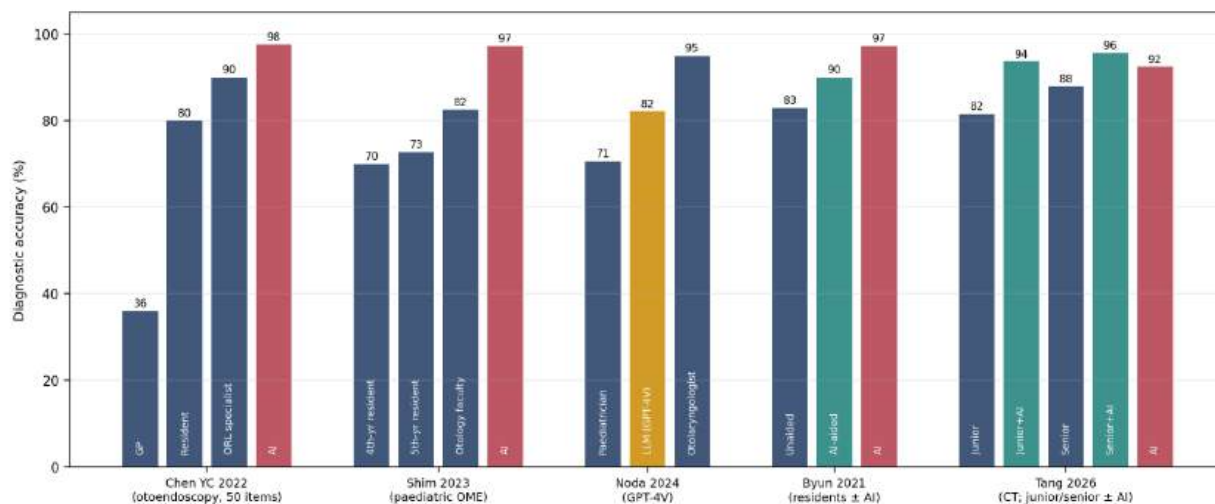


Figure 21. Diagnostic accuracy of AI compared with clinicians of different levels of experience, and the effect of AI assistance (study-level data).

**Table 10. Studies comparing AI with clinicians and evaluating AI-assisted interpretation**

| Study                          | Task                                | Clinician comparator                  | AI accuracy, % | Clinician accuracy, % | Difference, points | With AI assistance, % |
|--------------------------------|-------------------------------------|---------------------------------------|----------------|-----------------------|--------------------|-----------------------|
| Livingstone 2020 <sup>47</sup> | Otoscopy, 14 diagnoses              | Paediatric, EM, ORL and FM physicians | 88.7           | 58.9                  | +29.8              |                       |
| Wang YM 2020 <sup>73</sup>     | Temporal bone CT, 3 classes         | 2 otologists, 3 ORL, 1 radiologist    | 76.7           | 73.8                  | +2.9               |                       |
| Byun 2021 <sup>52</sup>        | Otoendoscopy, 4 classes             | 10 ORL residents                      | 97.2           | 82.9                  | +14.3              | 82.9 → 90.0           |
| Chen YC 2022 <sup>22</sup>     | Otoendoscopy, 10 classes (50 items) | GPs / residents / specialists         | 97.6*          | 36.0 / 80.0 / 90.0    | —                  |                       |

|                                 |                                           |                                       |           |                  |                   |             |
|---------------------------------|-------------------------------------------|---------------------------------------|-----------|------------------|-------------------|-------------|
| Takahashi<br>2022 <sup>75</sup> | CT mastoid<br>extension                   | 15<br>otolaryngologists               | 81.1      | 73.4             | +7.7              |             |
| Crowson<br>2023 <sup>54</sup>   | Paediatric<br>effusion, 3<br>classes      | 39 clinicians                         | 95.5      | 65.0             | +30.5             |             |
| Shim<br>2023 <sup>59</sup>      | Paediatric<br>OME                         | 2 otology<br>faculty, 5<br>residents  | 97.2      | 74.7 (69–<br>84) | +22.5             |             |
| Suresh<br>2024 <sup>28</sup>    | Otitis media,<br>handheld<br>otoscopy     | Paediatricians                        | 90.6      | 59.4             | +31.2             |             |
| Zeng J<br>2022 <sup>56</sup>    | CHL<br>prediction<br>from otoscopy        | 3 otologists                          | 81*       | 16–39            | —                 |             |
| Noda<br>2024 <sup>44</sup>      | Otoscopy, 4<br>classes (GPT-<br>4V)       | Paediatricians /<br>otolaryngologists | 82.1      | 70.6 /<br>>95    | +11.5 / –13       |             |
| Yağcı<br>2026 <sup>45</sup>     | CT<br>cholesteatoma<br>(GPT-<br>5/Gemini) | Radiologist                           | 43.6–49.5 | 96.0             | –46.5 to<br>–52.4 |             |
| Tang<br>2026 <sup>78</sup>      | CT<br>OME/COM, 3<br>classes               | Junior and senior<br>clinicians       | 92.4      | 81.5 /<br>87.9   | —                 | 93.6 / 95.6 |

*\*AI accuracy not measured on the identical reader set; excluded from pooled risk differences. CHL, conductive hearing loss; EM, emergency medicine; FM, family medicine; GP, general practitioner; ORL, otorhinolaryngology.*

## Internal-to-external generalisability

The most granular evidence on generalisability came from a cross-cohort experiment using three independent otoscopic datasets from Turkey, Chile and the United States. Across four architectures, mean internal AUROC was 0.94, falling to 0.77 when the same models were applied to unseen cohorts (mean decrement 0.17); a model achieving perfect internal discrimination on one cohort reached only 0.62–0.67 on another. Counterfactual experiments on the same public datasets showed that models could achieve high internal and external AUROC using clinically irrelevant artefacts, and systematic optimisation of pre-processing, augmentation and architecture raised external accuracy from 53% to 83%.<sup>17,18,72</sup>

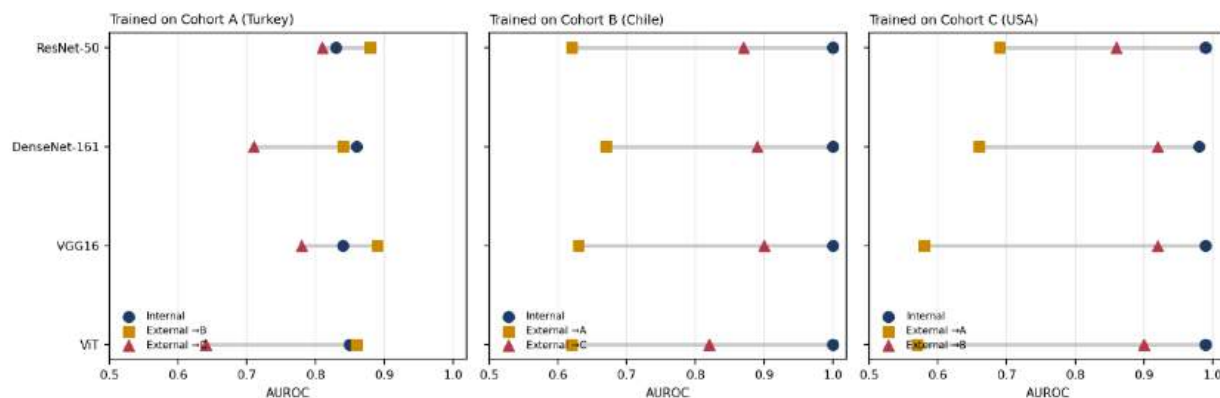


Figure 22. Internal versus external AUROC of four deep-learning architectures trained on each of three international otoscopic cohorts (data from Habib et al. 2023).

## Small-study effects, influence and robustness

The contour-enhanced funnel plot showed no material asymmetry (Egger intercept 0.28,  $P = 0.877$ ; Begg Kendall  $\tau = 0.152$ ,  $P = 0.191$ ), trim-and-fill imputed 0 studies, and the sample-size-based funnel plot was similarly symmetric. Because both tests have limited power and funnel asymmetry is not synonymous with publication bias, and because negative AI studies are less likely to be published, small-study effects cannot be excluded. The Galbraith plot identified 23 of 36 estimates outside the  $\pm 2$  band, confirming that heterogeneity is diffuse rather than driven by a few outliers, and the Baujat plot identified the very large audiogram dataset (Kassjański 2024) as the

dominant contributor to both heterogeneity and influence, followed by the two Zeng J studies, Chen BJ 2024 and Zeng X 2021.<sup>37,38,40–42,56,57</sup>

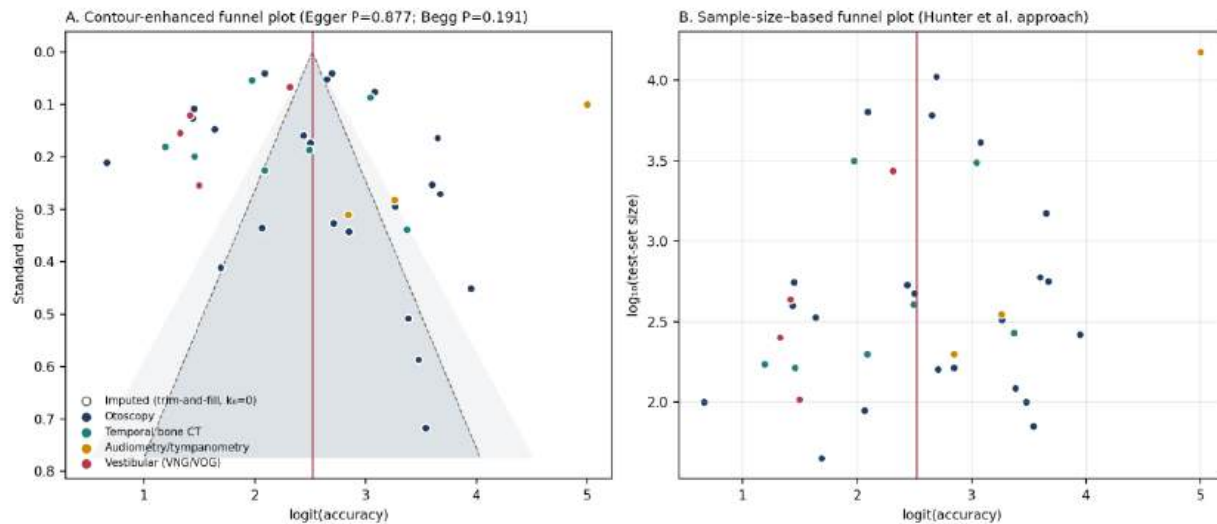


Figure 23. (A) Contour-enhanced funnel plot of logit accuracy against standard error with trim-and-fill; (B) sample-size-based funnel plot.

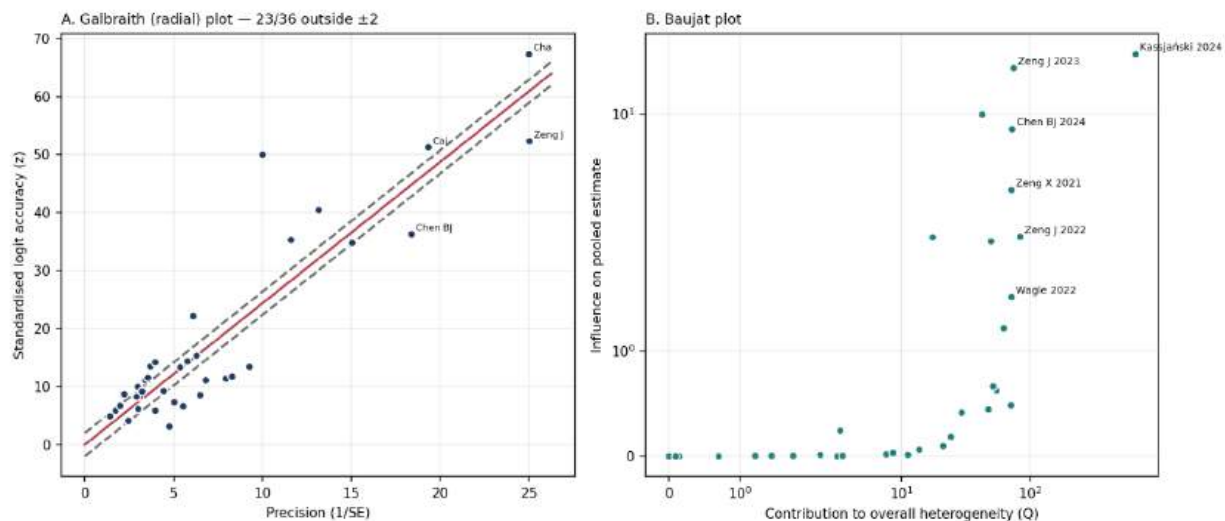


Figure 24. (A) Galbraith (radial) plot and (B) Baujat plot identifying contributors to heterogeneity and influence.

Leave-one-out analysis showed a stable pooled accuracy between 91.9% and 92.9%, and cumulative meta-analysis demonstrated that the pooled estimate stabilised above 90% from 2021 onward with progressively narrower confidence intervals. Across all prespecified sensitivity analyses the pooled accuracy remained between 89.0% and 92.9% ; the largest change occurred when the analysis was restricted to studies with surgical, histopathological or objective reference standards.<sup>32,33</sup>

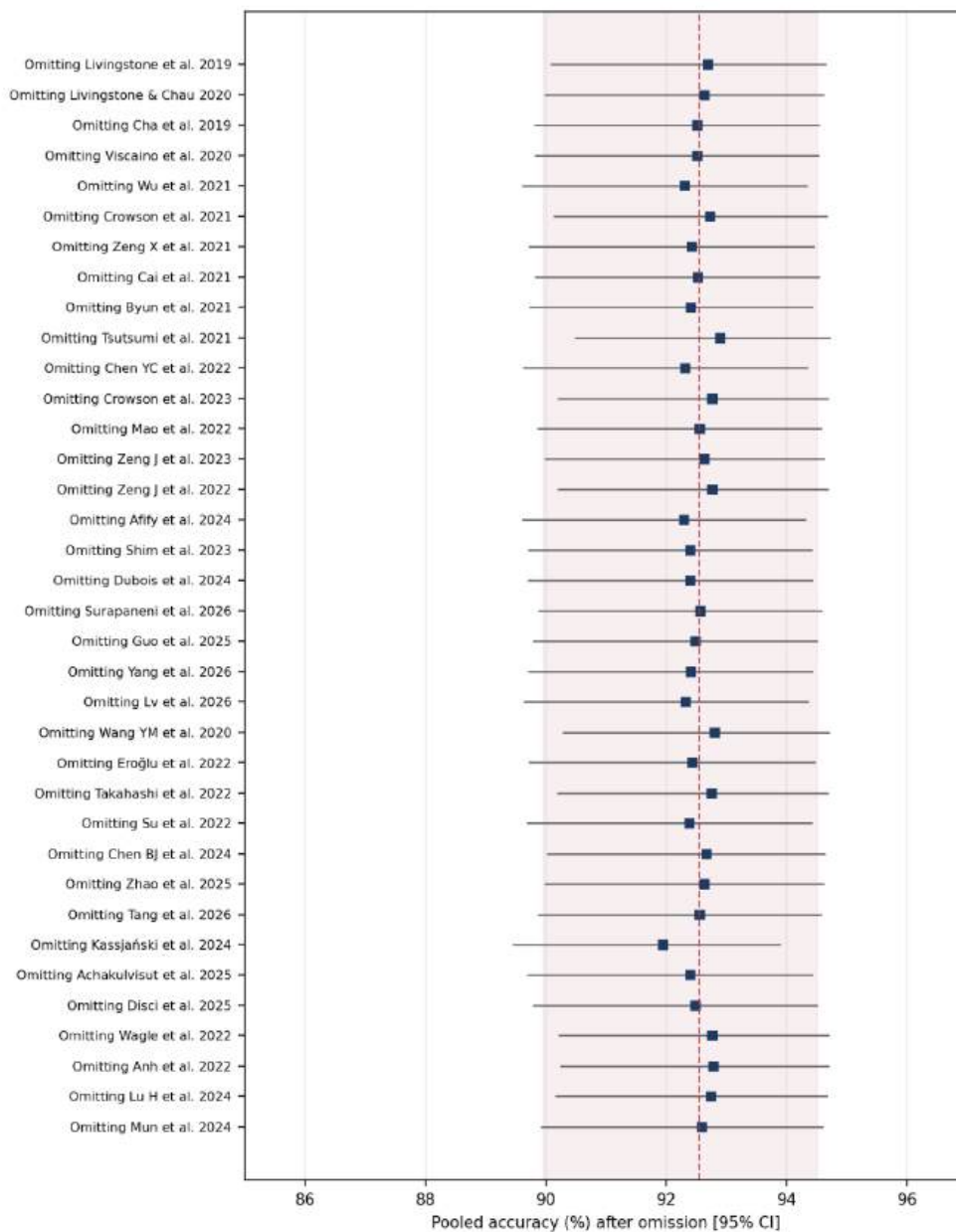


Figure 25. Leave-one-out sensitivity analysis of pooled accuracy.

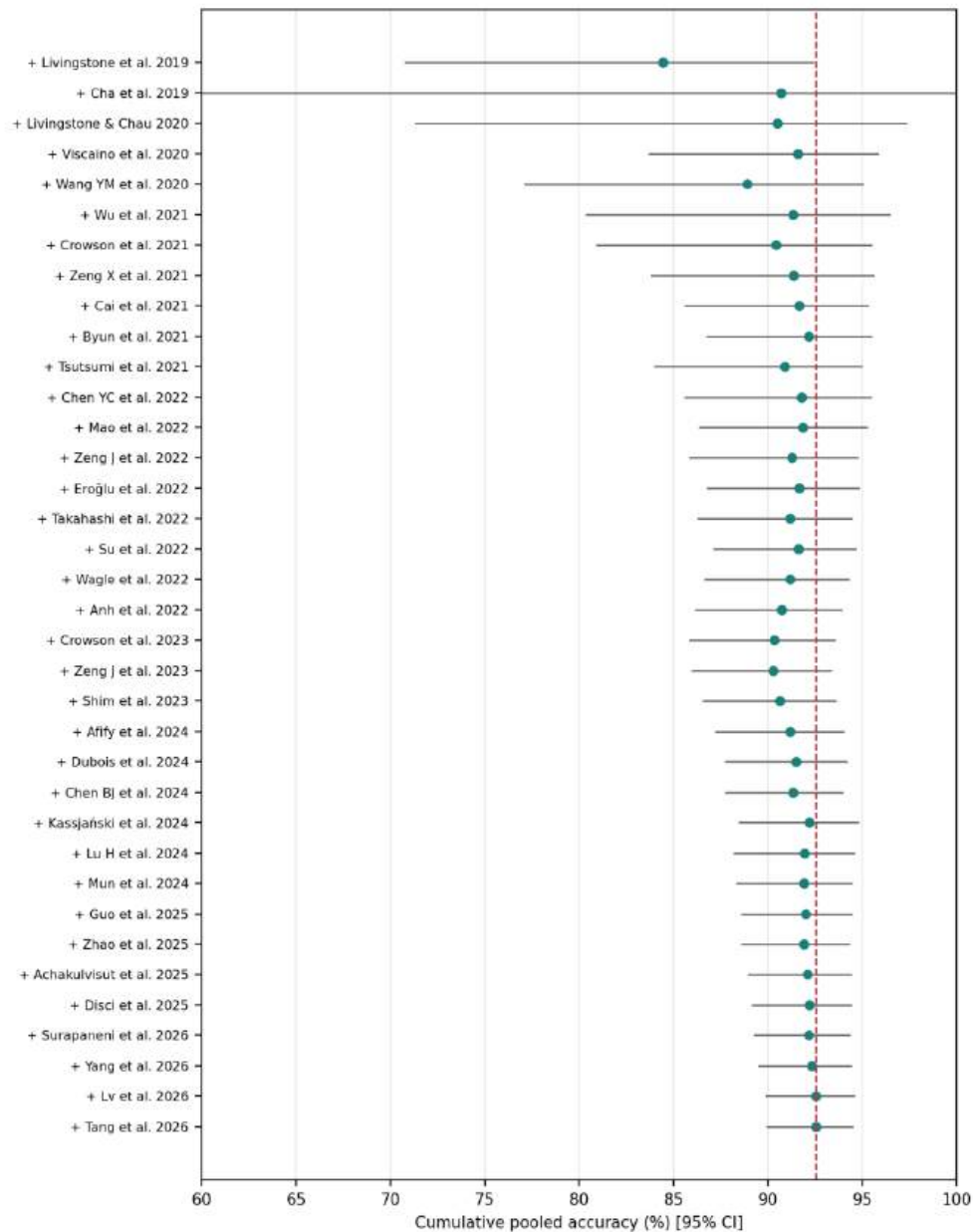


Figure 26. Cumulative meta-analysis of pooled accuracy by publication year.

**Table 11. Sensitivity analyses and small-study effect assessment for pooled accuracy**

| Analysis                                | k  | Accuracy, % (95% CI) | I <sup>2</sup> , % |
|-----------------------------------------|----|----------------------|--------------------|
| Primary analysis (REML + modified HKSJ) | 36 | 92.5 (90.0–94.5)     | 97.8               |

|                                                                         |        |                                                            |               |
|-------------------------------------------------------------------------|--------|------------------------------------------------------------|---------------|
| DerSimonian–Laird, Wald-type CI                                         | 36     | 92.5 (90.5–94.0)                                           | 97.8          |
| Common-effect (inverse-variance)                                        | 36     | 91.9 (91.7–92.2)                                           | 97.8          |
| Excluding cross-validation estimates                                    | 27     | 92.9 (90.1–95.0)                                           | 95.7          |
| Excluding approximated/reconstructed denominators                       | 31     | 92.5 (89.6–94.7)                                           | 98.0          |
| Excluding public/web-derived image datasets                             | 33     | 92.7 (90.2–94.7)                                           | 97.9          |
| Excluding test sets >5000 units                                         | 32     | 91.9 (89.2–94.0)                                           | 95.9          |
| Restricted to surgical/objective reference standard                     | 10     | 89.0 (83.4–92.8)                                           | 96.0          |
| Excluding clip/image-level unit-of-analysis studies (Mun, Eroğlu, Zhao) | 33     | 92.6 (89.7–94.7)                                           | 97.9          |
| Trim-and-fill (Duval–Tweedie L0, left side)                             | 36 + 0 | 92.5 (90.0–94.5)                                           | 97.8          |
| Leave-one-out range                                                     | 35     | 91.9–92.9                                                  | 96.3–<br>97.9 |
| Egger regression / Begg rank correlation                                | 36     | Intercept 0.28 (P = 0.877)<br>/ $\tau$ = 0.152 (P = 0.191) | —             |

### Translational readiness, opportunities and barriers

Mapping seven translational attributes across the 47 meta-analysed studies revealed a marked gap between technical performance and implementation readiness. Only 17.0% reported external or cross-site validation, 14.9% used prospectively collected data, 31.9% used surgical or objective reference standards, 21.3% provided explainability outputs, 36.2% compared AI with clinicians, 25.5% reported a mobile, web or edge deployment, and 21.3% studied paediatric cohorts; the median study satisfied only one of seven attributes, and only seven studies satisfied four or more.<sup>26,28,41,49,60,63</sup>

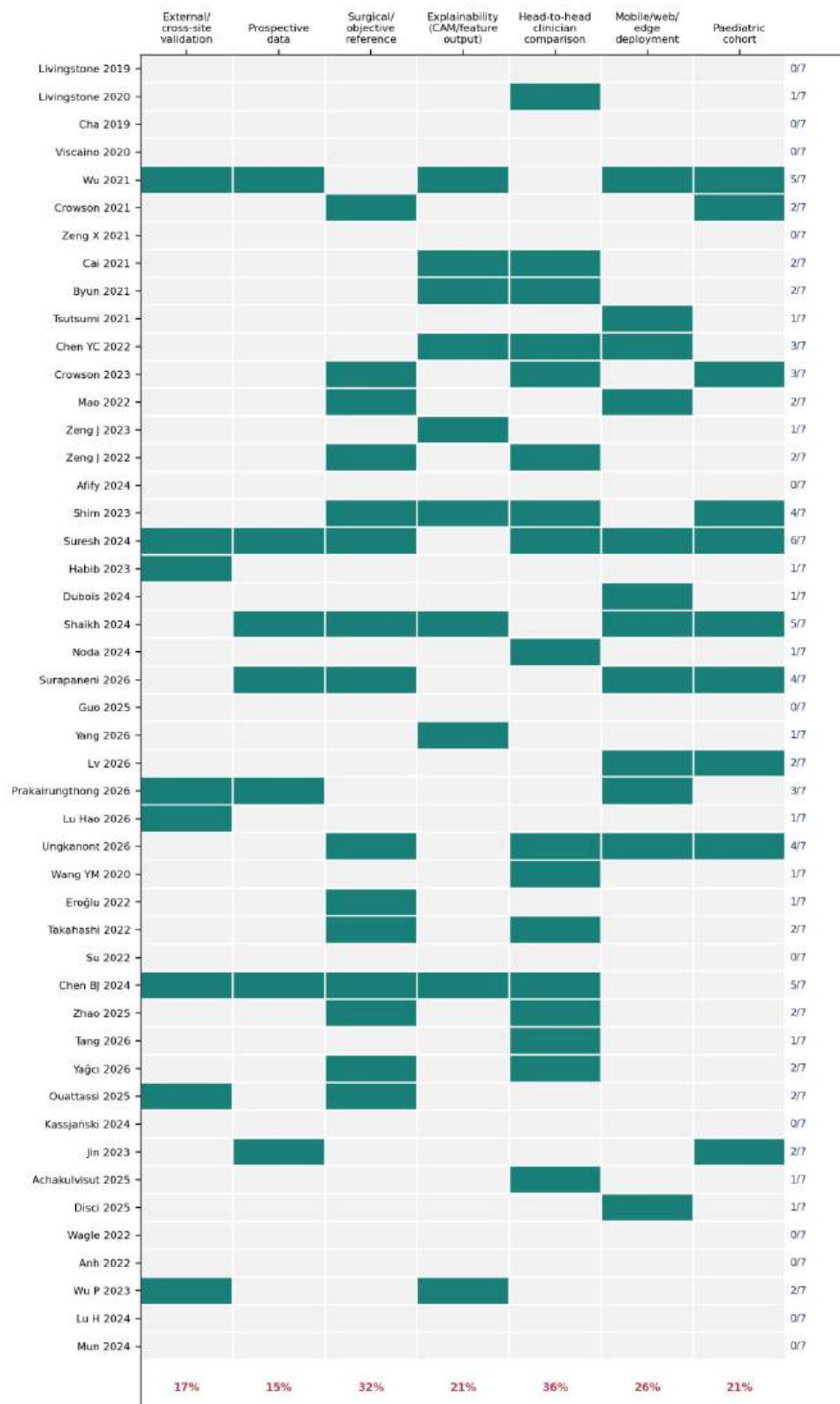


Figure 27. Translational readiness heat map of meta-analysed studies across seven implementation-relevant attributes (teal = reported). Row totals are shown on the right and column percentages at the bottom.

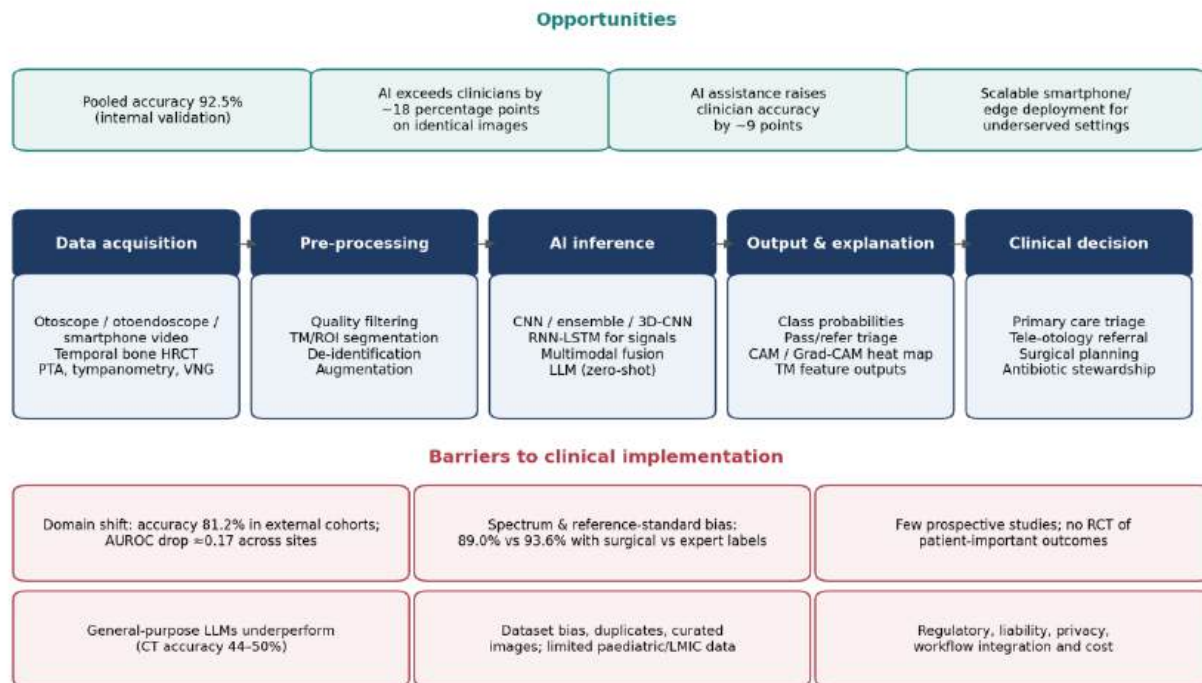


Figure 28. Conceptual framework of AI-enabled diagnostic otology: the diagnostic pipeline, evidence-based opportunities and barriers to clinical implementation identified in this review.

**Table 12. Evidence-based opportunities and barriers to clinical implementation of AI in diagnostic otology**

| Theme                                 | Key evidence from this review                                                                          | Implication for implementation                                    |
|---------------------------------------|--------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------|
| Opportunity: high diagnostic accuracy | Pooled accuracy 92.5%; sensitivity 92.6%; AUROC 0.957 under internal validation <sup>26,39,49,60</sup> | Suitable for triage/second-reader roles once externally validated |
| Opportunity: augmenting non-          | AI exceeded GPs/paediatricians by ~30 points; AI assistance +9.4 points <sup>28,47,52,78</sup>         | Greatest value in primary care, emergency care and                |

|                                                 |                                                                                                                                        |                                                                                 |
|-------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| specialists                                     |                                                                                                                                        | tele-otology                                                                    |
| Opportunity: point-of-care and edge deployment  | Smartphone apps, consumer-grade otoscopes, INT8 microcontroller deployment with preserved accuracy <sup>22,26,27,63</sup>              | Feasible in low-resource and archipelagic settings without cloud dependence     |
| Opportunity: multimodal and objective tasks     | Audiogram/tympanogram classification $\geq 94\%$ ; video + tympanometry fusion improved accuracy by 18 points <sup>42,65,85</sup>      | Integrate physiological and imaging data; enable layperson-led screening        |
| Barrier: domain shift                           | External/prospective accuracy 81.2%; mean AUROC drop 0.17 across cohorts; app accuracy 52.8% externally <sup>17,64,65</sup>            | Mandatory multisite external validation and local calibration before deployment |
| Barrier: reference-standard and spectrum bias   | Surgical/objective references 89.0% vs expert annotation 93.6%; 55% high-risk patient selection <sup>27,59,74</sup>                    | Use surgical/myringotomy confirmation and consecutive sampling                  |
| Barrier: dataset bias and leakage               | Dataset artefacts sufficient for high AUROC; duplicates in large datasets; image-level splits <sup>18,26,74</sup>                      | Patient-level partitioning, de-duplication and dataset documentation            |
| Barrier: limited prospective and trial evidence | Two prospective internal studies; no RCT of an AI diagnostic tool; RCTs of digital otoscopy alone show modest effects <sup>5,6,8</sup> | Pragmatic RCTs with antibiotic use, referral and hearing outcomes               |
| Barrier: general-purpose LLM reliability        | LLM accuracy 60.3%; CT accuracy 44–50%; poor reproducibility <sup>44,45</sup>                                                          | Not suitable as autonomous readers; require task-specific validation            |
| Barrier: governance                             | Few studies addressed explainability                                                                                                   | Regulatory approval,                                                            |

|              |                                                         |                                                               |
|--------------|---------------------------------------------------------|---------------------------------------------------------------|
| and workflow | (21%), integration, cost or regulation <sup>20,24</sup> | liability frameworks, EHR integration and economic evaluation |
|--------------|---------------------------------------------------------|---------------------------------------------------------------|

### Certainty of evidence

Certainty of evidence was rated low for pooled accuracy, sensitivity and the AI–clinician comparison, and very low for specificity, external-validation accuracy, AI-assisted interpretation and LLM performance, principally because of high risk of bias in patient selection and index-test evaluation, extreme inconsistency, and imprecision of prediction intervals.<sup>20,29</sup>

**Table 13. Summary of findings and GRADE certainty of evidence**

| Outcome                                   | Studies (k) | Pooled estimate (95% CI) | Risk of bias | Inconsistency                                 | Imprecision | Certainty        |
|-------------------------------------------|-------------|--------------------------|--------------|-----------------------------------------------|-------------|------------------|
| Diagnostic accuracy (internal validation) | 36          | 92.5% (90.0–94.5)        | Serious      | Serious (I <sup>2</sup> 97.8%; PI 64.7–98.8%) | Not serious | ⊕⊕○○<br>Low      |
| Sensitivity                               | 7           | 92.6% (89.4–94.8)        | Serious      | Not serious (I <sup>2</sup> 67.8%)            | Not serious | ⊕⊕○○<br>Low      |
| Specificity                               | 7           | 89.5% (74.3–96.1)        | Serious      | Very serious (I <sup>2</sup> 99.6%)           | Serious     | ⊕○○○<br>Very low |
| Accuracy in external/prospective cohorts  | 7           | 81.2% (59.1–92.8)        | Serious      | Very serious                                  | Serious     | ⊕○○○<br>Very low |
| AI vs clinicians (risk difference)        | 7           | +18.2 points             | Serious      | Serious (I <sup>2</sup> 73.1%)                | Not serious | ⊕⊕○○<br>Low      |

|                                      |                      |                           |             |              |              |                  |
|--------------------------------------|----------------------|---------------------------|-------------|--------------|--------------|------------------|
|                                      |                      | (7.3 to 29.0)             |             |              |              |                  |
| AI-assisted vs unassisted clinicians | 3 strata (2 studies) | +9.4 points (1.4 to 17.4) | Serious     | Not serious  | Very serious | ⊕○○○<br>Very low |
| General-purpose LLM accuracy         | 3 (2 studies)        | 60.3% (12.0–94.4)         | Not serious | Very serious | Very serious | ⊕○○○<br>Very low |

*Diagnostic-accuracy evidence from observational cross-sectional designs starts at high certainty under GRADE for diagnostic tests and was downgraded for risk of bias, inconsistency, indirectness and imprecision as indicated.*

## DISCUSSION

### Principal findings

This systematic review and meta-analysis synthesised 64 primary studies spanning the full scope of diagnostic otology. Four principal findings emerge. First, task-specific AI achieved a high pooled accuracy of 92.5% under internal validation, with sensitivity of 92.6% and an AUROC of 0.957, and performance was consistent across otoscopy, temporal bone CT and audiological interpretation, with vestibular eye-movement analysis lagging. Second, this performance did not transfer intact beyond development data: accuracy fell to 81.2% in external, prospective and point-of-care cohorts, and the prediction interval for a new setting spanned roughly 65% to 99%. Third, on identical cases, task-specific AI outperformed clinicians by approximately 18 percentage points—most markedly non-specialists—and improved clinician accuracy when used as an assistant, whereas general-purpose multimodal LLMs did not reach specialist performance. Fourth, the evidence base is methodologically immature: almost three-quarters of studies carried a high risk of bias, reference standards were frequently unverifiable, prospective data were rare and no randomised trial has evaluated an AI diagnostic tool in otology.<sup>17,39,41,44,45</sup>

### Comparison with previous systematic reviews

Our otoscopy estimate (93.2%) is consistent with earlier syntheses. Habib and colleagues reported 90.7% accuracy for normal-versus-abnormal classification and AI superiority over human assessors (93.4% vs 73.2%) in three studies, and Cao and colleagues reported pooled sensitivity and specificity of 93% and 85% with an AUC of 0.94 for tympanic membrane images, with otoendoscopic images outperforming handheld otoscopy. Song and colleagues found an average accuracy of 86% across 26 image-classification studies. The present review extends these findings by adding temporal bone CT, audiometry/tympanometry, vestibular testing and LLMs; by quantifying human comparisons and AI assistance as pooled risk differences; and by demonstrating, for the first time in otology, a statistically significant internal-to-external performance gap and a reference-standard effect.<sup>12–14</sup>

These patterns mirror the wider medical-imaging literature. Liu and colleagues found that only a minority of deep-learning studies performed out-of-sample external validation, and that reported equivalence with health-care professionals was undermined by poor reporting; Aggarwal and colleagues documented extreme heterogeneity across 503 studies; and a systematic review of radiology algorithms found that 81% showed decreased performance on external data, with a quarter declining by at least 0.10. The AUROC decrement of 0.17 observed across international otoscopic cohorts places otology at the less favourable end of this spectrum, plausibly because otoscopic images are acquired with heterogeneous devices, illumination, magnification, cerumen burden and operator technique.<sup>9,10,16,17</sup>

**Table 14. Comparison of the present review with previous systematic reviews and meta-analyses (cited in the Discussion only)**

| Review                   | Scope                                | Studies | Main findings pooled               | Additions of the present review |
|--------------------------|--------------------------------------|---------|------------------------------------|---------------------------------|
| Habib 2022 <sup>12</sup> | AI classification of otoscopy images | 39      | Normal vs abnormal accuracy 90.7%; | Four modalities; external       |

|                               |                                                   |                     |                                                                              |                                                                      |
|-------------------------------|---------------------------------------------------|---------------------|------------------------------------------------------------------------------|----------------------------------------------------------------------|
|                               |                                                   |                     | multiclass 97.6%; AI 93.4% vs humans 73.2%                                   | validation gap; reader and assistance effects as RD                  |
| Cao 2023 <sup>13</sup>        | Machine learning on TM images                     | 16                  | Sensitivity 93%; specificity 85%; AUC 0.94                                   | Reference-standard and task-complexity moderators; LLMs; GRADE       |
| Song 2023 <sup>14</sup>       | Image-based AI for middle ear disease (narrative) | 32                  | Average accuracy 86% (classification); 90.8% (segmentation + classification) | Quantitative pooling with REML/HKSJ and prediction intervals         |
| Chaturvedi 2026 <sup>15</sup> | Machine learning in videonystagmography for BPV   | Narrative synthesis | Methods and datasets; no quantitative pooling                                | Pooled vestibular accuracy (84.0%) within a cross-modality framework |

### Interpretation by modality

**Otoscopy and otoendoscopy.** The strongest and most mature evidence concerns tympanic membrane image classification. Large single-institution otoendoscopic archives yielded accuracies of 93–98%, and the largest studies demonstrated that performance scales with data volume. Clinically, the most relevant tasks are the distinction of AOM from OME and from a normal ear, which determines antibiotic prescribing; here a video-based classifier trained on smartphone recordings of predominantly young children achieved sensitivity and specificity of approximately

94%, and its feature outputs—particularly bulging—mirrored guideline diagnostic criteria, offering inherent clinical interpretability. Conversely, otoendoscopic accuracy overstates what can be expected from handheld otoscopy in uncooperative children, which explains the lower accuracies of operative and point-of-care paediatric studies anchored to myringotomy findings.<sup>27,39,40,49,50,59,60</sup>

Functional inference from images is an emerging frontier: a CNN predicted conductive hearing loss in OME from otoscopic images with 81% accuracy, far exceeding otologists who could not reliably infer the air–bone gap visually, while detection of attic retraction pockets and atelectasis showed high sensitivity but limited specificity. These tasks illustrate that AI can extract information beyond human perceptual limits but also that specificity, rather than sensitivity, is the weaker operating characteristic and must be calibrated to the clinical consequence of false positives.<sup>56,57</sup>

**Temporal bone CT.** CT-based models achieved 90.2% pooled accuracy for distinguishing normal ears, COM and cholesteatoma, with three-dimensional CNNs outperforming two-dimensional slice-based approaches and rivalling senior otologists. Importantly, CT studies most often used surgical or histopathological confirmation, so their slightly lower pooled accuracy probably reflects more rigorous reference standards rather than inferior technology. The explainable three-dimensional model that maintained an AUROC of 0.93 in an external institution and contributed to preoperative decision-making in 90.1% of prospectively evaluated patients represents the most implementation-ready otological AI identified, and the reader study in which junior clinicians improved from 81.5% to 93.6% accuracy with AI support demonstrates the potential to compress the experience curve for COM imaging.<sup>41,73,75,78</sup>

**Audiometry and tympanometry.** Classification of audiogram type and degree is a well-structured, rule-based problem, which explains accuracies above 94% and the narrow internal error. The principal clinical value lies not in replacing audiologists but in enabling task-shifting: a hybrid deep-learning tympanogram classifier matched audiologist performance on layperson-acquired tracings in rural Alaskan children, supporting community-based screening for preventable infection-related hearing loss. Integration of tympanometry with otoscopic video improved accuracy by 18

percentage points in an external cohort, underscoring the value of physiological–anatomical multimodal fusion.<sup>42,65,85–87</sup>

**Vestibular testing.** Nystagmus detection and BPPV subtype classification achieved the lowest pooled accuracy (84.0%), consistent with the technical challenge of torsional nystagmus, blinking artefacts, goggle slippage and low signal-to-noise recordings, and with single-expert labelling. Nevertheless, a composite one-dimensional deep-learning model achieved canal-specific AUROCs of 0.93–0.99 in an independent test cohort, and multimodal models integrating head position with eye movement markedly outperformed eye-only models (81.7% vs 52.8%), indicating that diagnostic context is essential for positional vertigo.<sup>43,90,92,94,95</sup>

### **Why does performance fall outside development data?**

Three mechanisms explain the internal-to-external gap. The first is domain shift: differences in otoscope or endoscope type, resolution, colour calibration, cropping and patient demographics alter the image distribution, and in the international cross-cohort experiment a model with perfect internal discrimination fell to an AUROC of 0.62 on a different cohort. The second is shortcut learning: counterfactual experiments showed that models trained on public otoscopy datasets could reach an AUROC of 0.89 using clinically irrelevant artefacts alone, and large datasets frequently contain near-duplicate images that inflate internal estimates. The third is spectrum and selection bias: curated, high-quality, enriched test sets exclude the blurred, cerumen-obscured and ambiguous images that dominate real-world primary care. Encouragingly, methodical optimisation of pre-processing, augmentation and architecture improved external accuracy from 53% to 83%, indicating that generalisability is partly an engineering problem amenable to solution.<sup>17,18,26,64,72</sup>

### **Reference-standard bias and the meaning of “accuracy”**

The finding that studies using surgical, myringotomy, histopathological or audiometric references reported lower accuracy than those using expert annotation (89.0% vs 93.6%) has important implications. When the reference standard is a clinician’s reading of the same image, AI is trained and tested to reproduce expert perception—including its errors—so agreement is inflated and the ceiling is set by expert accuracy, which for OME may be considerably lower than for perforation or cerumen. Anchoring labels to intraoperative findings, as in operative paediatric

596

cohorts and cholesteatoma CT studies, provides a more valid but more demanding benchmark. Future studies should prefer such references, report inter-rater agreement for expert labels and avoid evaluating AI against the same experts who generated the training labels.<sup>4,27,50,59,74,77</sup>

### **AI as a partner rather than a replacement**

The human-comparison evidence supports a collaborative model. The magnitude of AI superiority was inversely related to clinician expertise: approximately 30 percentage points over general physicians and paediatricians, approximately 14–26 points over residents, and only 3–8 points over otologists reading CT. AI assistance increased clinician accuracy by 9.4 points, comparable with the relative increases of 7% in sensitivity and 3% in specificity reported for deep learning–assisted clinicians in image-based cancer diagnostics. Clinicians’ ability to reject incorrect network advice when confident suggests that appropriately designed decision support can reduce automation bias. These data argue for deploying otological AI primarily as triage and second-reader support for non-specialists, where the absolute benefit is greatest, rather than as an autonomous replacement for otolaryngologists.<sup>11,22,28,47,52,78</sup>

### **General-purpose large language models**

Multimodal LLMs represent a qualitatively different class of AI: they are not trained on curated otological datasets but are accessible to any clinician with a smartphone. Our analysis shows that, despite outperforming paediatricians on otoscopic images when supplied with clinical data, these models performed at near-chance level on temporal bone CT, combined high sensitivity with very low specificity, and produced inconsistent answers to identical inputs one week apart. Their current role should be limited to education and hypothesis generation, and any clinical use requires the same rigorous, task-specific validation demanded of conventional diagnostic AI.<sup>44,45</sup>

### **Opportunities for clinical implementation**

Several opportunities are evident. Smartphone-attached otoscopes, consumer-grade digital otoscopes and quantised models running on low-cost microcontrollers make offline, point-of-care AI feasible in primary care, community screening and tele-otology, including in remote Indigenous communities where an algorithm achieved 77.8–99.3% accuracy across conditions. For Indonesia and other archipelagic, low-specialist-density settings, AI-supported triage at puskesmas level could

prioritise referral of chronic suppurative otitis media and suspected cholesteatoma, support guideline-concordant antibiotic decisions in AOM and OME, and enable layperson-led hearing screening linked to automated tympanogram interpretation. Explainability outputs—class activation maps highlighting perforation margins, retraction pockets or mastoid involvement—can also serve educational purposes for trainees.<sup>21,26,27,41,52,63,85</sup>

### **Barriers to clinical implementation**

Technical barriers include domain shift, dataset bias, class imbalance and the scarcity of curated paediatric and low- and middle-income-country datasets; the predominance of East Asian, single-centre archives in the included evidence raises concerns about performance across populations, devices and disease spectra. Methodological barriers include the reliance on cross-validation or validation sets used for model selection, image-level rather than patient-level partitioning, and the absence of prospective, consecutive-enrolment studies with prespecified operating thresholds. Clinical barriers include the need for adequate tympanic membrane visualisation and cerumen removal, the difficulty of image capture in uncooperative young children and the unresolved question of whether AI changes management. Systemic barriers comprise regulatory approval, medicolegal liability, data protection, integration into electronic health records and tele-otology workflows, reimbursement and cost-effectiveness—issues that the included studies scarcely addressed.<sup>4,17,18,20,29</sup>

Most importantly, diagnostic accuracy is a surrogate. Randomised trials of digital otoscopy without AI have produced modest gains in diagnostic agreement and no consistent reduction in antibiotic prescribing, whereas a pragmatic cluster trial of an optical coherence tomography otoscope reported halved odds of antibiotic prescribing at interim analysis. These precedents indicate that the value of otological AI must ultimately be demonstrated in pragmatic trials measuring antibiotic use, referral appropriateness, time to definitive treatment, hearing outcomes, complications and costs.<sup>5–8</sup>

### **Strengths**

Strengths of this review include its comprehensive cross-modality scope; restriction to recent evidence (2019–2026 within a 2014–2026 search window); verification of extracted

denominators against full texts; contemporary statistical methods (logit transformation, REML, modified HKSJ inference and prediction intervals) that avoid the known pitfalls of double-arcsine back-transformation and DerSimonian–Laird estimation; AI-adapted risk-of-bias assessment; extensive subgroup, meta-regression, influence and sensitivity analyses; and GRADE-based certainty ratings.<sup>20,30,32,33</sup>

### **Implications for research**

Future studies should (i) evaluate locked models on consecutive, prospectively enrolled patients in the intended-use setting, with handheld devices representative of primary care; (ii) use surgical, myringotomy, histopathological or audiometric reference standards, or adjudicated multi-expert consensus with reported agreement; (iii) partition data at the patient level, de-duplicate images and document dataset provenance; (iv) report full confusion matrices, calibration and operating thresholds aligned with STARD-AI, TRIPOD+AI and CONSORT-AI principles; (v) perform multisite external validation across devices, ethnicities and age groups, including Southeast Asian and Indonesian populations; (vi) conduct multi-reader multi-case studies quantifying AI assistance for general practitioners and nurses; and (vii) undertake pragmatic randomised trials and economic evaluations measuring patient-important and health-system outcomes. Adoption of the forthcoming QUADAS-AI and QUADAS-3 tools will improve the appraisal of future evidence.<sup>9,20,24,25,29</sup>

---

## **CONCLUSION AND RECOMMENDATIONS**

---

### **Conclusion**

Task-specific artificial intelligence achieves high diagnostic accuracy across diagnostic otology—otoscopy, temporal bone CT, audiometry/tympanometry and vestibular testing—with a pooled accuracy of 92.5% under internal validation, consistently outperforming non-specialist clinicians and improving clinician accuracy when used as an assistant. However, accuracy declines to approximately 81% in external, prospective and point-of-care settings, studies with rigorous surgical or objective reference standards report lower accuracy, general-purpose multimodal LLMs remain unreliable, and the certainty of evidence is low to very low. AI should currently be regarded

as a promising triage and decision-support adjunct rather than an autonomous diagnostic tool.<sup>17,41,45,78</sup>

### Recommendations

**For clinicians:** use validated, task-specific AI tools as a second reader or triage aid—particularly in primary care, emergency departments and tele-otology—while retaining clinical judgement, ensuring adequate tympanic membrane visualisation and confirming high-stakes diagnoses (cholesteatoma, persistent OME with hearing loss) by otological examination and appropriate investigations. General-purpose LLMs should not be used for diagnostic decisions.<sup>44,45,52,78</sup>

**For health systems and policymakers:** prioritise prospective, locally calibrated pilot deployments of smartphone- or edge-based otological AI within referral pathways in under-served regions, coupled with training, governance, data-protection safeguards and outcome monitoring.<sup>21,26,63</sup>

**For researchers and developers:** move beyond retrospective internal validation towards multicentre external validation, surgically confirmed reference standards, transparent reporting and pragmatic randomised trials of patient-important outcomes, with explicit attention to paediatric, Southeast Asian and low-resource populations.<sup>8,17,20</sup>

---

### DECLARATIONS

---

**Author contributions:** A.L.Z.T. conceptualised the review, performed the literature search, study selection, data extraction and risk-of-bias assessment, conducted the statistical analysis and drafted the manuscript. H.R.P.L. supervised the review, verified study selection, data extraction and clinical interpretation, and critically revised the manuscript. Both authors approved the final version.

**Funding:** This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

**Conflicts of interest:** The authors declare no conflicts of interest.

**Ethical approval:** Not required; this study synthesised published, de-identified aggregate data.

**Data availability:** The extraction dataset, risk-of-bias judgements and analysis code are available from the corresponding author on reasonable request.

---

## REFERENCES

---

1. Haile LM, Kamenov K, Briant PS, Orji AU, Steinmetz JD, Abdoli A, et al. Hearing loss prevalence and years lived with disability, 1990–2019: findings from the Global Burden of Disease Study 2019. *The Lancet*. 2021;397(10278):996-1009. doi: [10.1016/s0140-6736\(21\)00516-x](https://doi.org/10.1016/s0140-6736(21)00516-x). **Consensus:**  
[https://consensus.app/papers/details/5f932d41a59d510d8eba80a7ddc8eb34/?utm\\_source](https://consensus.app/papers/details/5f932d41a59d510d8eba80a7ddc8eb34/?utm_source)
2. McDaid D, Park AL, Chadha S. Estimating the global costs of hearing loss. *International Journal of Audiology*. 2021;60(3):162-70. doi: [10.1080/14992027.2021.1883197](https://doi.org/10.1080/14992027.2021.1883197). **Consensus:**  
[https://consensus.app/papers/details/ab6d8a3fa89f5a6baee6d0b796ee1963/?utm\\_source](https://consensus.app/papers/details/ab6d8a3fa89f5a6baee6d0b796ee1963/?utm_source)
3. Guo Z, Ji W, Song P, Zhao J, Yan M, Zou X, et al. Global, regional, and national burden of hearing loss in children and adolescents, 1990–2021: a systematic analysis from the Global Burden of Disease Study 2021. *BMC Public Health*. 2024;24(1):2521. doi: [10.1186/s12889-024-20010-0](https://doi.org/10.1186/s12889-024-20010-0). **Consensus:**  
[https://consensus.app/papers/details/d41ed0dc61625e5f874a3cb4ce66e878/?utm\\_source](https://consensus.app/papers/details/d41ed0dc61625e5f874a3cb4ce66e878/?utm_source)
4. Pichichero ME. Can Machine Learning and AI Replace Otoscopy for Diagnosis of Otitis Media? *Pediatrics*. 2021;147(4):e2020049584. doi: [10.1542/peds.2020-049584](https://doi.org/10.1542/peds.2020-049584). **Consensus:**  
[https://consensus.app/papers/details/15a034601bfd5430891466433c26c13a/?utm\\_source](https://consensus.app/papers/details/15a034601bfd5430891466433c26c13a/?utm_source)
5. Mousseau S, Lapointe A, Gravel J. Diagnosing acute otitis media using a smartphone otoscope; a randomized controlled trial. *The American Journal of Emergency Medicine*. 2018;36(10):1796-801. doi: [10.1016/j.ajem.2018.01.093](https://doi.org/10.1016/j.ajem.2018.01.093). **Consensus:**  
[https://consensus.app/papers/details/0acc6b851b1d5b4297dba7fe9cc9be69/?utm\\_source](https://consensus.app/papers/details/0acc6b851b1d5b4297dba7fe9cc9be69/?utm_source)
6. Kleinman K, Psoter KJ, Nyhan A, Solomon BS, Kim JM, Canares T. Evaluation of digital otoscopy in pediatric patients: A prospective randomized controlled clinical trial. *The American Journal of Emergency Medicine*. 2021;46:150-5. doi: [10.1016/j.ajem.2021.04.030](https://doi.org/10.1016/j.ajem.2021.04.030). **Consensus:**  
[https://consensus.app/papers/details/c767e54b4ae35e72ac86c51bc3e30643/?utm\\_source](https://consensus.app/papers/details/c767e54b4ae35e72ac86c51bc3e30643/?utm_source)
7. Chan KN, Silverstein A, Bryan LN, McCracken CE, Little WK, Shane AL. Comparison of a Smartphone Otoscope and Conventional Otoscope in the Diagnosis and Management of Acute

Otitis Media. Clin Pediatr (Phila). 2019;58(3):302-6. doi: [10.1177/0009922818812480](https://doi.org/10.1177/0009922818812480).

**Consensus:**

[https://consensus.app/papers/details/9eec414455425fd5bd71d47034569ca5/?utm\\_source](https://consensus.app/papers/details/9eec414455425fd5bd71d47034569ca5/?utm_source)

8. Zambelli-Weiner A, Nathanson BH, Calcagno F, Ansley J, Vattana K, Burnette-Vick B. Clinical utility of an optical coherence tomography middle ear scope: Interim results of the modification of antibiotic treatment intervention in children (OTO-MATIC) pragmatic cluster randomized controlled trial (RCT). International Journal of Pediatric Otorhinolaryngology. 2025;193:112356. doi: [10.1016/j.ijporl.2025.112356](https://doi.org/10.1016/j.ijporl.2025.112356). **Consensus:**  
[https://consensus.app/papers/details/ed6a3d9d965952268496eaa5e96ab7ca/?utm\\_source](https://consensus.app/papers/details/ed6a3d9d965952268496eaa5e96ab7ca/?utm_source)
9. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health. 2019;1(6):e271-97. doi: [10.1016/s2589-7500\(19\)30123-2](https://doi.org/10.1016/s2589-7500(19)30123-2). **Consensus:**  
[https://consensus.app/papers/details/32ec2621593e5347b7c997695034df13/?utm\\_source](https://consensus.app/papers/details/32ec2621593e5347b7c997695034df13/?utm_source)
10. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. npj Digit Med. 2021;4(1):65. doi: [10.1038/s41746-021-00438-z](https://doi.org/10.1038/s41746-021-00438-z). **Consensus:**  
[https://consensus.app/papers/details/633c78022253574793eb0fb7f91871d5/?utm\\_source](https://consensus.app/papers/details/633c78022253574793eb0fb7f91871d5/?utm_source)
11. Xue P, Si M, Qin D, Wei B, Seery S, Ye Z, et al. Unassisted Clinicians Versus Deep Learning–Assisted Clinicians in Image-Based Cancer Diagnostics: Systematic Review With Meta-analysis. J Med Internet Res. 2023;25:e43832. doi: [10.2196/43832](https://doi.org/10.2196/43832). **Consensus:**  
[https://consensus.app/papers/details/c239e528ea875fd1af35c9d322fd789e/?utm\\_source](https://consensus.app/papers/details/c239e528ea875fd1af35c9d322fd789e/?utm_source)
12. Habib A, Kajbafzadeh M, Hasan Z, Wong E, Gunasekera H, Perry C, et al. Artificial intelligence to classify ear disease from otoscopy: A systematic review and meta-analysis. Clinical Otolaryngology. 2022;47(3):401-13. doi: [10.1111/coa.13925](https://doi.org/10.1111/coa.13925). **Consensus:**  
[https://consensus.app/papers/details/3d71b053830c5553ac5496b3e7e6a24c/?utm\\_source](https://consensus.app/papers/details/3d71b053830c5553ac5496b3e7e6a24c/?utm_source)
13. Cao Z, Chen F, Grais EM, Yue F, Cai Y, Swanepoel DW, et al. Machine Learning in Diagnosing Middle Ear Disorders Using Tympanic Membrane Images: A Meta-Analysis. The

- Laryngoscope. 2023;133(4):732-41. doi: [10.1002/lary.30291](https://doi.org/10.1002/lary.30291). **Consensus:**  
[https://consensus.app/papers/details/22e1a12d68d0514280f44bb7a983a699/?utm\\_source](https://consensus.app/papers/details/22e1a12d68d0514280f44bb7a983a699/?utm_source)
14. Song D, Kim T, Lee Y, Kim J. Image-Based Artificial Intelligence Technology for Diagnosing Middle Ear Diseases: A Systematic Review. JCM. 2023;12(18):5831. doi: [10.3390/jcm12185831](https://doi.org/10.3390/jcm12185831). **Consensus:**  
[https://consensus.app/papers/details/38d8d352553a5a118c0c2eeb38bdecc5/?utm\\_source](https://consensus.app/papers/details/38d8d352553a5a118c0c2eeb38bdecc5/?utm_source)
15. Chaturvedi K, Yang N, Dansereau D, Lovejoy C, Braytee A, Welgampola M, et al. Diagnosis of Benign Positional Vertigo: A Systematic Review of Machine Learning and Deep Learning within Videonystagmography. ACM Comput Surv. 2026;58(8):1-29. doi: [10.1145/3789494](https://doi.org/10.1145/3789494). **Consensus:**  
[https://consensus.app/papers/details/645e9bacf223576598cde317e429fe48/?utm\\_source](https://consensus.app/papers/details/645e9bacf223576598cde317e429fe48/?utm_source)
16. Yu AC, Mohajer B, Eng J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. Radiology: Artificial Intelligence. 2022;4(3):e210064. doi: [10.1148/ryai.210064](https://doi.org/10.1148/ryai.210064). **Consensus:**  
[https://consensus.app/papers/details/a2aac152d6e858969ef6b2083c64bea1/?utm\\_source](https://consensus.app/papers/details/a2aac152d6e858969ef6b2083c64bea1/?utm_source)
17. Habib AR, Xu Y, Bock K, Mohanty S, Sederholm T, Weeks WB, et al. Evaluating the generalizability of deep learning image classification algorithms to detect middle ear disease using otoscopy. Sci Rep. 2023;13(1):5368. doi: [10.1038/s41598-023-31921-0](https://doi.org/10.1038/s41598-023-31921-0). **Consensus:**  
[https://consensus.app/papers/details/0e57126a5ee45e198bd91170ddcefe14/?utm\\_source](https://consensus.app/papers/details/0e57126a5ee45e198bd91170ddcefe14/?utm_source)
18. Xu Y, Habib AR, Crossland G, Patel H, Perry C, Bock K, et al. Towards reliable use of artificial intelligence to classify otitis media using otoscopic images: Addressing bias and improving data quality. PLoS One. 2026;21(5):e0338867. doi: [10.1371/journal.pone.0338867](https://doi.org/10.1371/journal.pone.0338867). **Consensus:**  
[https://consensus.app/papers/details/a5c5109868e35200906a43fb3f423c90/?utm\\_source](https://consensus.app/papers/details/a5c5109868e35200906a43fb3f423c90/?utm_source)
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev. 2021;10(1):89. doi: [10.1186/s13643-021-01626-4](https://doi.org/10.1186/s13643-021-01626-4). **Consensus:**  
[https://consensus.app/papers/details/c1a1d4c752e857f2ba6639ed813f181a/?utm\\_source](https://consensus.app/papers/details/c1a1d4c752e857f2ba6639ed813f181a/?utm_source)

20. Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med.* 2021;27(10):1663-5. doi: [10.1038/s41591-021-01517-0](https://doi.org/10.1038/s41591-021-01517-0). **Consensus:** [https://consensus.app/papers/details/ead02a079daf5456958b04318ce185ca/?utm\\_source](https://consensus.app/papers/details/ead02a079daf5456958b04318ce185ca/?utm_source)
21. Habib AR, Crossland G, Patel H, Wong E, Kong K, Gunasekera H, et al. An Artificial Intelligence Computer-vision Algorithm to Triage Otoscopic Images From Australian Aboriginal and Torres Strait Islander Children. *Otology & Neurotology.* 2022;43(4):481-8. doi: [10.1097/mao.00000000000003484](https://doi.org/10.1097/mao.00000000000003484). **Consensus:** [https://consensus.app/papers/details/524f156500cc544cb0ad9fdc3783eeb0/?utm\\_source](https://consensus.app/papers/details/524f156500cc544cb0ad9fdc3783eeb0/?utm_source)
22. Chen YC, Chu YC, Huang CY, Lee YT, Lee WY, Hsu CY, et al. Smartphone-based artificial intelligence using a transfer learning algorithm for the detection and diagnosis of middle ear diseases: A retrospective deep learning study. *eClinicalMedicine.* 2022;51:101543. doi: [10.1016/j.eclinm.2022.101543](https://doi.org/10.1016/j.eclinm.2022.101543). **Consensus:** [https://consensus.app/papers/details/8a30e0f8f14255aebafbfd9038ad6d16/?utm\\_source](https://consensus.app/papers/details/8a30e0f8f14255aebafbfd9038ad6d16/?utm_source)
23. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ.* 2021;372:n160. doi: [10.1136/bmj.n160](https://doi.org/10.1136/bmj.n160). **Consensus:** [https://consensus.app/papers/details/1ce155fe5ccc5127a671f21cc92d59b6/?utm\\_source](https://consensus.app/papers/details/1ce155fe5ccc5127a671f21cc92d59b6/?utm_source)
24. Guni A, Sounderajah V, Whiting P, Bossuyt P, Darzi A, Ashrafian H. Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies Using AI (QUADAS-AI): Protocol for a Qualitative Study. *JMIR Res Protoc.* 2024;13:e58202. doi: [10.2196/58202](https://doi.org/10.2196/58202). **Consensus:** [https://consensus.app/papers/details/7719d184ec13554297a11c6fc451a224/?utm\\_source](https://consensus.app/papers/details/7719d184ec13554297a11c6fc451a224/?utm_source)
25. Whiting PF, Tomlinson E, Rutjes AWS, Davenport CF, Yang B, Westwood ME, et al. QUADAS-3: A Revised Tool for the Quality Assessment of Diagnostic Test Accuracy Studies. *Ann Intern Med.* 2026;179(4):548-55. doi: [10.7326/annals-25-02104](https://doi.org/10.7326/annals-25-02104). **Consensus:** [https://consensus.app/papers/details/7f7199c86eae57f78ec25e6ebe2997bb/?utm\\_source](https://consensus.app/papers/details/7f7199c86eae57f78ec25e6ebe2997bb/?utm_source)
26. Dubois C, Eigen D, Simon F, Couloigner V, Gormish M, Chalumeau M, et al. Development and validation of a smartphone-based deep-learning-enabled system to detect middle-ear conditions

in otoscopic images. npj Digit Med. 2024;7(1):162. doi: [10.1038/s41746-024-01159-9](https://doi.org/10.1038/s41746-024-01159-9).

**Consensus:**

[https://consensus.app/papers/details/31a184c8bf0356c0a2715a2e43e8966f/?utm\\_source](https://consensus.app/papers/details/31a184c8bf0356c0a2715a2e43e8966f/?utm_source)

27. Surapaneni S, Rangarajan N, Davis K, Pletcher K, Flowers J, Strub G, et al. Artificial intelligence classification of pediatric middle ear effusion using consumer-grade otoscopes. International Journal of Pediatric Otorhinolaryngology. 2026;200:112655. doi: [10.1016/j.ijporl.2025.112655](https://doi.org/10.1016/j.ijporl.2025.112655).

**Consensus:**

[https://consensus.app/papers/details/617693b64fbe5018a259baed0d3fb3ed/?utm\\_source](https://consensus.app/papers/details/617693b64fbe5018a259baed0d3fb3ed/?utm_source)

28. Suresh K, Wu MP, Benboujja F, Christakis B, Newton A, Hartnick CJ, et al. AI Model Versus Clinician Otoscopy in the Operative Setting for Otitis Media Diagnosis. Otolaryngol--head neck surg. 2024;170(6):1598-601. doi: [10.1002/ohn.559](https://doi.org/10.1002/ohn.559).

**Consensus:**

[https://consensus.app/papers/details/4dc0b85217d350f2a130cc0e28e83cca/?utm\\_source](https://consensus.app/papers/details/4dc0b85217d350f2a130cc0e28e83cca/?utm_source)

29. Jayakumar S, Sounderajah V, Normahani P, Harling L, Markar SR, Ashrafian H, et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. npj Digit Med. 2022;5(1):11. doi: [10.1038/s41746-021-00544-y](https://doi.org/10.1038/s41746-021-00544-y).

**Consensus:**

[https://consensus.app/papers/details/6584fef2fad05d87a656d33b33fdffc4/?utm\\_source](https://consensus.app/papers/details/6584fef2fad05d87a656d33b33fdffc4/?utm_source)

30. Schwarzer G, Chemaitelly H, Abu-Raddad LJ, Rücker G. Seriously misleading results using inverse of Freeman-Tukey double arcsine transformation in meta-analysis of single proportions. Research Synthesis Methods. 2019;10(3):476-83. doi: [10.1002/jrsm.1348](https://doi.org/10.1002/jrsm.1348).

**Consensus:**

[https://consensus.app/papers/details/cc835eb79aa4587e9f4f37e10b5a92b8/?utm\\_source](https://consensus.app/papers/details/cc835eb79aa4587e9f4f37e10b5a92b8/?utm_source)

31. Lin L, Chu H. Meta-analysis of Proportions Using Generalized Linear Mixed Models. Epidemiology. 2020;31(5):713-7. doi: [10.1097/ede.0000000000001232](https://doi.org/10.1097/ede.0000000000001232).

**Consensus:**

[https://consensus.app/papers/details/fed169035fa95442a0f9d86a20770924/?utm\\_source](https://consensus.app/papers/details/fed169035fa95442a0f9d86a20770924/?utm_source)

32. Langan D, Higgins JPT, Jackson D, Bowden J, Veroniki AA, Kontopantelis E, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Research Synthesis Methods. 2019;10(1):83-98. doi: [10.1002/jrsm.1316](https://doi.org/10.1002/jrsm.1316).

**Consensus:**

[https://consensus.app/papers/details/13b576d97fe75ca28bbef621c8bf3995/?utm\\_source](https://consensus.app/papers/details/13b576d97fe75ca28bbef621c8bf3995/?utm_source)

33. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol*. 2014;14(1):25. doi: [10.1186/1471-2288-14-25](https://doi.org/10.1186/1471-2288-14-25). **Consensus:**  
[https://consensus.app/papers/details/cb0f8813c7555c329bde150d9ac3e847/?utm\\_source](https://consensus.app/papers/details/cb0f8813c7555c329bde150d9ac3e847/?utm_source)
34. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Med Res Methodol*. 2015;15(1):99. doi: [10.1186/s12874-015-0091-1](https://doi.org/10.1186/s12874-015-0091-1). **Consensus:**  
[https://consensus.app/papers/details/dc41b84af1b45432aa6ae01d5100bcdf/?utm\\_source](https://consensus.app/papers/details/dc41b84af1b45432aa6ae01d5100bcdf/?utm_source)
35. Veroniki AA, Jackson D, Bender R, Kuss O, Langan D, Higgins JPT, et al. Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. *Research Synthesis Methods*. 2019;10(1):23-43. doi: [10.1002/jrsm.1319](https://doi.org/10.1002/jrsm.1319). **Consensus:**  
[https://consensus.app/papers/details/45ee8b7c41b25c3eab659dbcc08adaf7/?utm\\_source](https://consensus.app/papers/details/45ee8b7c41b25c3eab659dbcc08adaf7/?utm_source)
36. Nyaga VN, Arbyn M, Aerts M. Metaprop: a Stata command to perform meta-analysis of binomial data. *Arch Public Health*. 2014;72(1):39. doi: [10.1186/2049-3258-72-39](https://doi.org/10.1186/2049-3258-72-39). **Consensus:**  
[https://consensus.app/papers/details/5d3702f17b45521cbdd9d1471a6e394a/?utm\\_source](https://consensus.app/papers/details/5d3702f17b45521cbdd9d1471a6e394a/?utm_source)
37. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*. 2011;343:d4002. doi: [10.1136/bmj.d4002](https://doi.org/10.1136/bmj.d4002). **Consensus:**  
[https://consensus.app/papers/details/37fc12e758405694a8068a87d7f3a59d/?utm\\_source](https://consensus.app/papers/details/37fc12e758405694a8068a87d7f3a59d/?utm_source)
38. Hunter JP, Saratzis A, Sutton AJ, Boucher RH, Sayers RD, Bown MJ. In meta-analyses of proportion studies, funnel plots were found to be an inaccurate method of assessing publication bias. *Journal of Clinical Epidemiology*. 2014;67(8):897-903. doi: [10.1016/j.jclinepi.2014.03.003](https://doi.org/10.1016/j.jclinepi.2014.03.003). **Consensus:**  
[https://consensus.app/papers/details/ef8ef266903e5e548abe75c33e0f9c73/?utm\\_source](https://consensus.app/papers/details/ef8ef266903e5e548abe75c33e0f9c73/?utm_source)
39. Cha D, Pae C, Seong SB, Choi JY, Park HJ. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. *EBioMedicine*. 2019;45:606-14. doi:

[10.1016/j.ebiom.2019.06.050](https://doi.org/10.1016/j.ebiom.2019.06.050).

**Consensus:**

[https://consensus.app/papers/details/15ef29175f1b5dac9a1ad310b7bc3a07/?utm\\_source](https://consensus.app/papers/details/15ef29175f1b5dac9a1ad310b7bc3a07/?utm_source)

40. Zeng X, Jiang Z, Luo W, Li H, Li H, Li G, et al. Efficient and accurate identification of ear diseases using an ensemble deep learning model. Sci Rep. 2021;11(1):10839. doi: [10.1038/s41598-021-90345-w](https://doi.org/10.1038/s41598-021-90345-w).

**Consensus:**

[https://consensus.app/papers/details/89de71bba2c552f68d7ba09e40ef3f26/?utm\\_source](https://consensus.app/papers/details/89de71bba2c552f68d7ba09e40ef3f26/?utm_source)

41. Chen B, Li Y, Sun Y, Sun H, Wang Y, Lyu J, et al. A 3D and Explainable Artificial Intelligence Model for Evaluation of Chronic Otitis Media Based on Temporal Bone Computed Tomography: Model Development, Validation, and Clinical Application. J Med Internet Res. 2024;26:e51706. doi: [10.2196/51706](https://doi.org/10.2196/51706).

**Consensus:**

[https://consensus.app/papers/details/6f19b051d53250afb4d67cab20024400/?utm\\_source](https://consensus.app/papers/details/6f19b051d53250afb4d67cab20024400/?utm_source)

42. Kassjański M, Kulawiak M, Przewoźny T, Tretiakow D, Kuryłowicz J, Molisz A, et al. Automated hearing loss type classification based on pure tone audiometry data. Sci Rep. 2024;14(1):14203. doi: [10.1038/s41598-024-64310-2](https://doi.org/10.1038/s41598-024-64310-2).

**Consensus:**

[https://consensus.app/papers/details/eb6741949e3752cfbc3241ce8d7fad41/?utm\\_source](https://consensus.app/papers/details/eb6741949e3752cfbc3241ce8d7fad41/?utm_source)

43. Wu P, Liu X, Dai Q, Yu J, Zhao J, Yu F, et al. Diagnosing the benign paroxysmal positional vertigo via 1D and deep-learning composite model. J Neurol. 2023;270(8):3800-9. doi: [10.1007/s00415-023-11662-w](https://doi.org/10.1007/s00415-023-11662-w).

**Consensus:**

[https://consensus.app/papers/details/6191bd1e769559e69f167b62f0d683b4/?utm\\_source](https://consensus.app/papers/details/6191bd1e769559e69f167b62f0d683b4/?utm_source)

44. Noda M, Yoshimura H, Okubo T, Kosu R, Uchiyama Y, Nomura A, et al. Feasibility of Multimodal Artificial Intelligence Using GPT-4 Vision for the Classification of Middle Ear Disease: Qualitative Study and Validation. JMIR AI. 2024;3:e58342. doi: [10.2196/58342](https://doi.org/10.2196/58342).

**Consensus:**

[https://consensus.app/papers/details/5deb1d959abc5583a3c5587bd63d0da4/?utm\\_source](https://consensus.app/papers/details/5deb1d959abc5583a3c5587bd63d0da4/?utm_source)

45. Yağcı B, Palaz S, Çetinkaya EA, Şenol U. Diagnostic performance and short-interval reproducibility of multimodal large language models in differentiating cholesteatoma from chronic otitis media using key-image temporal bone high-resolution computed tomography.

- 608

[10.3390/jcm10153198](https://doi.org/10.3390/jcm10153198).

**Consensus:**

[https://consensus.app/papers/details/33254da896205d2db34e39a0861e4c96/?utm\\_source](https://consensus.app/papers/details/33254da896205d2db34e39a0861e4c96/?utm_source)

53. Tsutsumi K, Goshtasbi K, Risbud A, Khosravi P, Pang JC, Lin HW, et al. A Web-Based Deep Learning Model for Automated Diagnosis of Otoscopic Images. *Otology & Neurotology*. 2021;42(9):e1382-8. doi: [10.1097/mao.00000000000003210](https://doi.org/10.1097/mao.00000000000003210). **Consensus:**

[https://consensus.app/papers/details/71fc7272c8bc5bb0a07b630c593becaf/?utm\\_source](https://consensus.app/papers/details/71fc7272c8bc5bb0a07b630c593becaf/?utm_source)

54. Crowson MG, Bates DW, Suresh K, Cohen MS, Hartnick CJ. “Human vs Machine” Validation of a Deep Learning Algorithm for Pediatric Middle Ear Infection Diagnosis. *Otolaryngol--head neck surg*. 2023;169(1):41-6. doi: [10.1177/01945998221119156](https://doi.org/10.1177/01945998221119156). **Consensus:**

[https://consensus.app/papers/details/e561001328ba598c933ae0459f9af938/?utm\\_source](https://consensus.app/papers/details/e561001328ba598c933ae0459f9af938/?utm_source)

55. Mao C, Li A, Hu J, Wang P, Peng D, Wang J, et al. Efficient and accurate diagnosis of otomycosis using an ensemble deep-learning model. *Front Mol Biosci*. 2022;9:951432. doi: [10.3389/fmolb.2022.951432](https://doi.org/10.3389/fmolb.2022.951432). **Consensus:**

[https://consensus.app/papers/details/9ebfbb8c3ae15edfa1a3a5657dfcd7/?utm\\_source](https://consensus.app/papers/details/9ebfbb8c3ae15edfa1a3a5657dfcd7/?utm_source)

56. Zeng J, Kang W, Chen S, Lin Y, Deng W, Wang Y, et al. A Deep Learning Approach to Predict Conductive Hearing Loss in Patients With Otitis Media With Effusion Using Otoscopic Images. *JAMA Otolaryngol Head Neck Surg*. 2022;148(7):612. doi: [10.1001/jamaoto.2022.0900](https://doi.org/10.1001/jamaoto.2022.0900). **Consensus:**

[https://consensus.app/papers/details/c9976d1a289555e3be214d1494f5e7f6/?utm\\_source](https://consensus.app/papers/details/c9976d1a289555e3be214d1494f5e7f6/?utm_source)

57. Zeng J, Deng W, Yu J, Xiao L, Chen S, Zhang X, et al. A deep learning approach to the diagnosis of atelectasis and attic retraction pocket in otitis media with effusion using otoscopic images. *Eur Arch Otorhinolaryngol*. 2023;280(4):1621-7. doi: [10.1007/s00405-022-07632-z](https://doi.org/10.1007/s00405-022-07632-z).

**Consensus:**

[https://consensus.app/papers/details/3af9cd7a33575b16bd3afd4b42d1eb8f/?utm\\_source](https://consensus.app/papers/details/3af9cd7a33575b16bd3afd4b42d1eb8f/?utm_source)

58. Afify HM, Mohammed KK, Hassanien AE. Insight into Automatic Image Diagnosis of Ear Conditions Based on Optimized Deep Learning Approach. *Ann Biomed Eng*. 2024;52(4):865-76. doi: [10.1007/s10439-023-03422-8](https://doi.org/10.1007/s10439-023-03422-8). **Consensus:**

[https://consensus.app/papers/details/87fdbd33d0115367ae6c59ca38d27334/?utm\\_source](https://consensus.app/papers/details/87fdbd33d0115367ae6c59ca38d27334/?utm_source)

59. Shim JH, Sunwoo W, Choi BY, Kim KG, Kim YJ. Improving the Accuracy of Otitis Media with Effusion Diagnosis in Pediatric Patients Using Deep Learning. *Bioengineering*. 2023;10(11):1337. doi: [10.3390/bioengineering10111337](https://doi.org/10.3390/bioengineering10111337). **Consensus:** [https://consensus.app/papers/details/faf40bb9a59059cfa1391c17d4d3b3b0/?utm\\_source](https://consensus.app/papers/details/faf40bb9a59059cfa1391c17d4d3b3b0/?utm_source)
60. Shaikh N, Conway SJ, Kovačević J, Condessa F, Shope TR, Haralam MA, et al. Development and Validation of an Automated Classifier to Diagnose Acute Otitis Media in Children. *JAMA Pediatr*. 2024;178(4):401. doi: [10.1001/jamapediatrics.2024.0011](https://doi.org/10.1001/jamapediatrics.2024.0011). **Consensus:** [https://consensus.app/papers/details/a81d3fe8c79757bc8a44ae27f361f283/?utm\\_source](https://consensus.app/papers/details/a81d3fe8c79757bc8a44ae27f361f283/?utm_source)
61. Guo Q, Xie L, Zhou L. Deep learning for otitis media classification using otoscopic image. *Medicine*. 2025;104(48):e46218. doi: [10.1097/md.00000000000046218](https://doi.org/10.1097/md.00000000000046218). **Consensus:** [https://consensus.app/papers/details/62d7b1883b445c7bb6f8889e15d44f98/?utm\\_source](https://consensus.app/papers/details/62d7b1883b445c7bb6f8889e15d44f98/?utm_source)
62. Yang CY, Lee CJ, Lai WS, Chen KY, Kuo CF, Liu CH, et al. Deep Learning-Based Semantic Segmentation and Classification of Otoscopic Images for Otitis Media Diagnosis and Health Promotion. *Diagnostics*. 2026;16(3):467. doi: [10.3390/diagnostics16030467](https://doi.org/10.3390/diagnostics16030467). **Consensus:** [https://consensus.app/papers/details/a67ad2e44a09593e94640179b922fd05/?utm\\_source](https://consensus.app/papers/details/a67ad2e44a09593e94640179b922fd05/?utm_source)
63. Lv C, Jia D, Gao B, Xia L, Wang X, Wu Z. Deep learning–assisted otoscopic screening for paediatric otitis media: feasibility of edge deployment. *Acta Oto-Laryngologica*. 2026;146(6):740-8. doi: [10.1080/00016489.2026.2632953](https://doi.org/10.1080/00016489.2026.2632953). **Consensus:** [https://consensus.app/papers/details/8d7d76729ff5564eafe51407db0a8c51/?utm\\_source](https://consensus.app/papers/details/8d7d76729ff5564eafe51407db0a8c51/?utm_source)
64. Prakairungthong S, Limviriyakul S, Thitisomboon W, Bhabhatsatam B, Thongyai K, Atipas S, et al. OTOLens: Development and validation of a mobile application for assessment of tympanic membrane and middle ear pathology. *American Journal of Otolaryngology*. 2026;47(6):104949. doi: [10.1016/j.amjoto.2026.104949](https://doi.org/10.1016/j.amjoto.2026.104949). **Consensus:** [https://consensus.app/papers/details/4703535893915653b6ea9989e26b2592/?utm\\_source](https://consensus.app/papers/details/4703535893915653b6ea9989e26b2592/?utm_source)
65. Lu H, Langefeld CD, Zinnia A, Demir MF, Chavan S, Elmaraghy CA, et al. Confidence-guided integration of otoscopy videos and tympanometry for middle ear disease diagnosis: A multi-center otoscopy study with external paired-cohort evaluation. *PLoS One*. 2026;21(9):e0357492.

- doi: [10.1371/journal.pone.0357492](https://doi.org/10.1371/journal.pone.0357492). **Consensus:**  
[https://consensus.app/papers/details/a2a2eda2bbba55d7ba0b2716bf2f959d/?utm\\_source](https://consensus.app/papers/details/a2a2eda2bbba55d7ba0b2716bf2f959d/?utm_source)
66. Ungkanont K, Udomchaiporn A, Sripchoong N, Wannarong T, Rugsujrit T, Chueprasert T, et al. An artificial intelligence model for the diagnosis of otitis media with effusion in children. DIGITAL HEALTH. 2026;12:20552076261450814. doi: [10.1177/20552076261450814](https://doi.org/10.1177/20552076261450814). **Consensus:**  
[https://consensus.app/papers/details/37a127ba70b156dfa686e4419519eee3/?utm\\_source](https://consensus.app/papers/details/37a127ba70b156dfa686e4419519eee3/?utm_source)
67. Sundgaard JV, Harte J, Bray P, Laugesen S, Kamide Y, Tanaka C, et al. Deep metric learning for otitis media classification. Medical Image Analysis. 2021;71:102034. doi: [10.1016/j.media.2021.102034](https://doi.org/10.1016/j.media.2021.102034). **Consensus:**  
[https://consensus.app/papers/details/3ca36fb401485d73a0b011fde421ad13/?utm\\_source](https://consensus.app/papers/details/3ca36fb401485d73a0b011fde421ad13/?utm_source)
68. Choi Y, Chae J, Park K, Hur J, Kweon J, Ahn JH. Automated multi-class classification for prediction of tympanic membrane changes with deep learning models. PLoS ONE. 2022;17(10):e0275846. doi: [10.1371/journal.pone.0275846](https://doi.org/10.1371/journal.pone.0275846). **Consensus:**  
[https://consensus.app/papers/details/546e15301ddf5ac3b4ddec1415ad97be/?utm\\_source](https://consensus.app/papers/details/546e15301ddf5ac3b4ddec1415ad97be/?utm_source)
69. Park YS, Jeon JH, Kong TH, Chung TY, Seo YJ. Deep Learning Techniques for Ear Diseases Based on Segmentation of the Normal Tympanic Membrane. Clin Exp Otorhinolaryngol. 2023;16(1):28-36. doi: [10.21053/ceo.2022.00675](https://doi.org/10.21053/ceo.2022.00675). **Consensus:**  
[https://consensus.app/papers/details/b3a87a4060485fe0ae1b52bccfc18117/?utm\\_source](https://consensus.app/papers/details/b3a87a4060485fe0ae1b52bccfc18117/?utm_source)
70. Chang KM, Surapaneni SS, Shaikh N, Marston AP, Vecchiotti MA, Rangarajan N, et al. Pediatric tympanostomy tube assessment via deep learning. American Journal of Otolaryngology. 2024;45(4):104334. doi: [10.1016/j.amjoto.2024.104334](https://doi.org/10.1016/j.amjoto.2024.104334). **Consensus:**  
[https://consensus.app/papers/details/11acc1e99f765bd5acbaecbda7e88693/?utm\\_source](https://consensus.app/papers/details/11acc1e99f765bd5acbaecbda7e88693/?utm_source)
71. Seo HW, Ko DW, Oh J, Lee J, Ji YB, Han SY, et al. Development and Validation of a CNN-Based Diagnostic Pipeline for the Diagnosis of Otitis Media. JCM. 2025;14(23):8572. doi: [10.3390/jcm14238572](https://doi.org/10.3390/jcm14238572). **Consensus:**  
[https://consensus.app/papers/details/bf391b7006b3507eb8f4d6d7a6f33e50/?utm\\_source](https://consensus.app/papers/details/bf391b7006b3507eb8f4d6d7a6f33e50/?utm_source)

72. Lindmark M, Lundberg T, Werner M, Soda P, Gronlund C, Ohberg F. Toward Generalizable Image Classification Models for Detecting Middle Ear Diseases. IEEE Access. 2026;14:131655-68. doi: [10.1109/access.2026.3723951](https://doi.org/10.1109/access.2026.3723951). **Consensus:** [https://consensus.app/papers/details/db3d7658255a53d4856fe9d4352f6bb8/?utm\\_source](https://consensus.app/papers/details/db3d7658255a53d4856fe9d4352f6bb8/?utm_source)
73. Wang YM, Li Y, Cheng YS, He ZY, Yang JM, Xu JH, et al. Deep Learning in Automated Region Proposal and Diagnosis of Chronic Otitis Media Based on Computed Tomography. Ear & Hearing. 2020;41(3):669-77. doi: [10.1097/aud.0000000000000794](https://doi.org/10.1097/aud.0000000000000794). **Consensus:** [https://consensus.app/papers/details/bde5701084d85698b88d1c264855f73d/?utm\\_source](https://consensus.app/papers/details/bde5701084d85698b88d1c264855f73d/?utm_source)
74. Eroğlu O, Eroğlu Y, Yıldırım M, Karlıdag T, Çınar A, Akyiğit A, et al. Is it useful to use computerized tomography image-based artificial intelligence modelling in the differential diagnosis of chronic otitis media with and without cholesteatoma? American Journal of Otolaryngology. 2022;43(3):103395. doi: [10.1016/j.amjoto.2022.103395](https://doi.org/10.1016/j.amjoto.2022.103395). **Consensus:** [https://consensus.app/papers/details/b0264153cc4a551cbb533b801c31319c/?utm\\_source](https://consensus.app/papers/details/b0264153cc4a551cbb533b801c31319c/?utm_source)
75. Takahashi M, Noda K, Yoshida K, Tsuchida K, Yui R, Nakazawa T, et al. Preoperative prediction by artificial intelligence for mastoid extension in pars flaccida cholesteatoma using temporal bone high-resolution computed tomography: A retrospective study. PLoS ONE. 2022;17(10):e0273915. doi: [10.1371/journal.pone.0273915](https://doi.org/10.1371/journal.pone.0273915). **Consensus:** [https://consensus.app/papers/details/8c417d90377c5dc9b0e8bcc1aeabd0d/?utm\\_source](https://consensus.app/papers/details/8c417d90377c5dc9b0e8bcc1aeabd0d/?utm_source)
76. Su R, Song J, Wang Z, Mao S, Mao Y, Wu X, et al. Application of high resolution computed tomography image assisted classification model of middle ear diseases based on 3D-convolutional neural network. Zhong Nan Da Xue Xue Bao Yi Xue Ban. 2022;47(8):1037-48. doi: [10.11817/j.issn.1672-7347.2022.210704](https://doi.org/10.11817/j.issn.1672-7347.2022.210704). **Consensus:** [https://consensus.app/papers/details/3c21d774b9255b9b92e3d45e058a77a8/?utm\\_source](https://consensus.app/papers/details/3c21d774b9255b9b92e3d45e058a77a8/?utm_source)
77. Zhao Y, Ma R, Ren H, Feng N, Zhang N, Wang L, et al. The computer-aided diagnosis model of middle ear cholesteatoma based on integrated convolutional neural networks. Zhonghua Er Bi Yan Hou Tou Jing Wai Ke Za Zhi. 2025;60(5):511-9. doi: [10.3760/cma.j.cn115330-20240515-00282](https://doi.org/10.3760/cma.j.cn115330-20240515-00282). **Consensus:** [https://consensus.app/papers/details/2ddac5f576f65c1db6b4ca7377c8f5f1/?utm\\_source](https://consensus.app/papers/details/2ddac5f576f65c1db6b4ca7377c8f5f1/?utm_source)

78. Tang Y, Fan W, Cheng Z, Leng Y, Liu Y, Wu T, et al. Deep learning for automated diagnosis and differentiation of otitis media on temporal bone CT. BMC Med Imaging. 2026;26(1):316. doi: [10.1186/s12880-026-02386-0](https://doi.org/10.1186/s12880-026-02386-0). **Consensus:** [https://consensus.app/papers/details/a6133a3e8a9c5b6c9fae3e5e89f1af20/?utm\\_source](https://consensus.app/papers/details/a6133a3e8a9c5b6c9fae3e5e89f1af20/?utm_source)
79. Ouattassi N, Zoizou A, Maaroufi M, Labiyad EM, Benatiya Andaloussi T, Benmansour N, et al. Evaluating Partial Middle Ear Filling for the Diagnosis of Cholesteatoma Using Machine Learning: A Radiological Approach. Cureus. 2025;17:e83503. doi: [10.7759/cureus.83503](https://doi.org/10.7759/cureus.83503). **Consensus:** [https://consensus.app/papers/details/e1288a4961da5fb7bd08a1a36996d818/?utm\\_source](https://consensus.app/papers/details/e1288a4961da5fb7bd08a1a36996d818/?utm_source)
80. Duan B, Pan LL, Chen WX, Qiao ZW, Xu ZM. An in-depth discussion of cholesteatoma, middle ear Inflammation, and langerhans cell histiocytosis of the temporal bone, based on diagnostic results. Front Pediatr. 2022;10:809523. doi: [10.3389/fped.2022.809523](https://doi.org/10.3389/fped.2022.809523). **Consensus:** [https://consensus.app/papers/details/62fa60875a6157bf9a30b387d68771e2/?utm\\_source](https://consensus.app/papers/details/62fa60875a6157bf9a30b387d68771e2/?utm_source)
81. Duan B, Guo Z, Pan L, Xu Z, Chen W. Temporal bone CT-based deep learning models for differential diagnosis of primary ciliary dyskinesia related otitis media and simple otitis media with effusion. Am J Transl Res. 2022;14(7):4728-35. **Consensus:** [https://consensus.app/papers/details/48f8a58bda585424b4bec65aab04e6d/?utm\\_source](https://consensus.app/papers/details/48f8a58bda585424b4bec65aab04e6d/?utm_source)
82. Cao C, Song J, Su R, Wu X, Wang Z, Hou M. Structure-constrained deep feature fusion for chronic otitis media and cholesteatoma identification. Multimed Tools Appl. 2023;82(29):45869-89. doi: [10.1007/s11042-023-15425-7](https://doi.org/10.1007/s11042-023-15425-7). **Consensus:** [https://consensus.app/papers/details/19c8d155694c5fc2865b010b31490303/?utm\\_source](https://consensus.app/papers/details/19c8d155694c5fc2865b010b31490303/?utm_source)
83. Wang Z, Song J, Su R, Hou M, Qi M, Zhang J, et al. Structure-aware deep learning for chronic middle ear disease. Expert Systems with Applications. 2022;194:116519. doi: [10.1016/j.eswa.2022.116519](https://doi.org/10.1016/j.eswa.2022.116519). **Consensus:** [https://consensus.app/papers/details/8c11cfdda0ef5f8c903386c8f2ca7cfb/?utm\\_source](https://consensus.app/papers/details/8c11cfdda0ef5f8c903386c8f2ca7cfb/?utm_source)
84. Qi F, You Z, Guo J, Hong Y, Wu X, Zhang D, et al. An automatic diagnosis model of otitis media with high accuracy rate using transfer learning. Front Mol Biosci. 2024;10:1250596. doi:

[10.3389/fmolb.2023.1250596](https://doi.org/10.3389/fmolb.2023.1250596).

**Consensus:**

[https://consensus.app/papers/details/a08c2cb891db5c7499d38e49814c6ac1/?utm\\_source](https://consensus.app/papers/details/a08c2cb891db5c7499d38e49814c6ac1/?utm_source)

85. Jin FQ, Huang O, Kleindienst Robler S, Morton S, Platt A, Egger JR, et al. A Hybrid Deep Learning Approach to Identify Preventable Childhood Hearing Loss. *Ear & Hearing*. 2023;44(5):1262-70. doi: [10.1097/aud.0000000000001380](https://doi.org/10.1097/aud.0000000000001380). **Consensus:**

[https://consensus.app/papers/details/d5a227b63e525c26807fca34e73048a5/?utm\\_source](https://consensus.app/papers/details/d5a227b63e525c26807fca34e73048a5/?utm_source)

86. Achakulvisut T, Phanthong S, Timpitak T, Vesessook K, Junthong S, Utainrat W, et al. Developing a diagnostic support system for audiogram interpretation using deep learning-based object detection. *Journal of Otology*. 2025;20(1):26-32. doi: [10.26599/joto.2025.9540005](https://doi.org/10.26599/joto.2025.9540005).

**Consensus:**

[https://consensus.app/papers/details/8e391327b91f5a3db43ff18b72742449/?utm\\_source](https://consensus.app/papers/details/8e391327b91f5a3db43ff18b72742449/?utm_source)

87. Disci AY, Konukseven O, Tanisir Disci R. Comparison of two methods in determining hearing loss type and hearing loss degree: mobile application coded with artificial neural networks and conditional structures. *Egypt J Otolaryngol*. 2025;41(1):103. doi: [10.1186/s43163-025-00849-9](https://doi.org/10.1186/s43163-025-00849-9).

**Consensus:**

[https://consensus.app/papers/details/011973abb0e15350972b2306a1a076f8/?utm\\_source](https://consensus.app/papers/details/011973abb0e15350972b2306a1a076f8/?utm_source)

88. Crowson MG, Lee JW, Hamour A, Mahmood R, Babier A, Lin V, et al. AutoAudio: Deep Learning for Automatic Audiogram Interpretation. *J Med Syst*. 2020;44(9):163. doi: [10.1007/s10916-020-01627-1](https://doi.org/10.1007/s10916-020-01627-1).

**Consensus:**

[https://consensus.app/papers/details/78f5003cd03a5618aee5736cc19ca6aa/?utm\\_source](https://consensus.app/papers/details/78f5003cd03a5618aee5736cc19ca6aa/?utm_source)

89. Dou Z, Li Y, Deng D, Zhang Y, Pang A, Fang C, et al. Pure tone audiogram classification using deep learning techniques. *Clinical Otolaryngology*. 2024;49(5):595-603. doi: [10.1111/coa.14170](https://doi.org/10.1111/coa.14170).

**Consensus:**

[https://consensus.app/papers/details/2dadf6546fb75e0f95506cef26d53ae4/?utm\\_source](https://consensus.app/papers/details/2dadf6546fb75e0f95506cef26d53ae4/?utm_source)

90. Wagle N, Morkos J, Liu J, Reith H, Greenstein J, Gong K, et al. aEYE: A deep learning system for video nystagmus detection. *Front Neurol*. 2022;13:963968. doi: [10.3389/fneur.2022.963968](https://doi.org/10.3389/fneur.2022.963968).

**Consensus:**

[https://consensus.app/papers/details/2ea4f71c628055f68950aaff37d79f22/?utm\\_source](https://consensus.app/papers/details/2ea4f71c628055f68950aaff37d79f22/?utm_source)

91. Anh DT, Takakura H, Asai M, Ueda N, Shojaku H. Application of machine learning in the diagnosis of vestibular disease. Sci Rep. 2022;12(1):20805. doi: [10.1038/s41598-022-24979-9](https://doi.org/10.1038/s41598-022-24979-9).  
**Consensus:**  
[https://consensus.app/papers/details/3cbb9166f3935a4aa2c13b6a770c45f0/?utm\\_source](https://consensus.app/papers/details/3cbb9166f3935a4aa2c13b6a770c45f0/?utm_source)
92. Lu H, Mao Y, Li J, Zhu L. Multimodal deep learning-based diagnostic model for BPPV. BMC Med Inform Decis Mak. 2024;24(1):82. doi: [10.1186/s12911-024-02438-x](https://doi.org/10.1186/s12911-024-02438-x). **Consensus:**  
[https://consensus.app/papers/details/469ddcb212a856cbb5795366e3777128/?utm\\_source](https://consensus.app/papers/details/469ddcb212a856cbb5795366e3777128/?utm_source)
93. Mun SB, Kim YJ, Lee JH, Han GC, Cho SH, Jin S, et al. Deep Learning-Based Nystagmus Detection for BPPV Diagnosis. Sensors. 2024;24(11):3417. doi: [10.3390/s24113417](https://doi.org/10.3390/s24113417).  
**Consensus:**  
[https://consensus.app/papers/details/e0cef5d3640554bca194ab18213d3419/?utm\\_source](https://consensus.app/papers/details/e0cef5d3640554bca194ab18213d3419/?utm_source)
94. Lim EC, Park JH, Jeon HJ, Kim HJ, Lee HJ, Song CG, et al. Developing a Diagnostic Decision Support System for Benign Paroxysmal Positional Vertigo Using a Deep-Learning Model. JCM. 2019;8(5):633. doi: [10.3390/jcm8050633](https://doi.org/10.3390/jcm8050633). **Consensus:**  
[https://consensus.app/papers/details/6a0fb73ca38253f083f4e8b295b86700/?utm\\_source](https://consensus.app/papers/details/6a0fb73ca38253f083f4e8b295b86700/?utm_source)
95. Kong S, Huang Z, Deng W, Zhan Y, Lv J, Cui Y. Nystagmus patterns classification framework based on deep learning and optical flow. Computers in Biology and Medicine. 2023;153:106473. doi: [10.1016/j.combiomed.2022.106473](https://doi.org/10.1016/j.combiomed.2022.106473). **Consensus:**  
[https://consensus.app/papers/details/8faf3fa156d655958ef286a90e3c8017/?utm\\_source](https://consensus.app/papers/details/8faf3fa156d655958ef286a90e3c8017/?utm_source)