frontiers in Audiology and Otology

TYPE Original Research
PUBLISHED 30 September 2026
DOI 10.3389/fauot.2026.1952229

Check for updates

OPEN ACCESS

EDITED BY
Neeraj Singh Thakur,
University of Oklahoma Health Sciences Center, United States

REVIEWED BY
Liam Barrett,
University College London, United Kingdom
Marta Campi,
Institut Pasteur, France

*CORRESPONDENCE
Seyed-Ali Sadegh-Zadeh
✉ ali.sadegh-zadeh@staffs.ac.uk

†These authors have contributed equally to this work and share first authorship

RECEIVED 29 July 2026
REVISED 29 August 2026
ACCEPTED 31 August 2026
PUBLISHED 30 September 2026

CITATION
Sadegh-Zadeh S-A, Saadat S, Barati MA, Sharvia S and Jarchi D (2026) Label construction can determine measured detector performance on small auditory brainstem response benchmarks. *Front. Audiol. Otol.* 4:1952229. doi: 10.3389/fauot.2026.1952229

COPYRIGHT
© 2026 Sadegh-Zadeh, Saadat, Barati, Sharvia and Jarchi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Label construction can determine measured detector performance on small auditory brainstem response benchmarks

Seyed-Ali Sadegh-Zadeh[1*†], Shayan Saadat[2†], Mohammad Amin Barati[3], Septavera Sharvia[4] and Delaram Jarchi[5]

[1]Department of Computing, University of Staffordshire, Stoke-on-Trent, United Kingdom, [2]Queen Elizabeth Hospital Birmingham, University Hospitals Birmingham, Birmingham, United Kingdom, [3]School of Mechanical Engineering, College of Engineering, University of Tehran, Tehran, Iran, [4]Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia, [5]Department of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom

Background: Two small public corpora, the PhysioNet earndb tone-pip set and the Mendeley click set, recur as evaluation data for objective auditory brainstem response (ABR) detection. Neither corpus includes expert annotations of response presence. Studies that cannot commission their own annotations therefore construct the label automatically, most commonly by declaring a response present when two repeated sub-averages correlate above a threshold—the clinical two-replication reproducibility criterion. Whether detector comparisons built on that automatic label would survive a switch to expert annotation has not been tested.

Methods: Audiologist annotations published by Liu et al. (2025) were mapped onto the recordings (one rater, 204 earndb conditions; six raters, 86 Mendeley conditions) and the mapping validated against the record inventories. On identical data, we constructed the reproducibility surrogate and four detectors spanning the capacity range (the instrument statistic wnsfmp, logistic regression, gradient boosting, a multilayer perceptron). Each learned detector was trained under each label and scored against each label (a 2 × 2 factorial) by leave-one-subject-out cross-validation; discrimination is reported pooled and as a frequency-adjusted AUROC; inference rests on exact subject-level permutation tests.

Results: The surrogate agreed with expert judgment at $\kappa = 0.167$ (earndb) and 0.248 (Mendeley), against $\kappa = 0.691$ among the six Mendeley audiologists. It is largely a function of signal-to-noise ratio: a single-replicate instrument statistic predicted it with AUROC 0.789 on sub-averages in which the expert saw no response, and a simulation with no response reproduced this. Training on the surrogate cost 0.139–0.304 pooled AUROC against expert judgment (perceptron: every informative listener, exact $p = 0.031$), a loss arising across listeners and frequencies rather than within them; scoring fixed predictions against either label did not reorder detectors. Changing both, as a study would, reordered them: wnsfmp fell from 0.827 to 0.662, while the perceptron rose from 0.713 to 0.898 (interaction −0.350, exact $p = 0.031$); about 37% of this is pooling across frequencies with different expert prevalence, and the within-frequency remainder was not resolved. Rescoring against individual Mendeley raters moved AUROC by 0.065–0.082, below the smallest difference eight listeners can resolve

(0.096–0.136). The stimulus level alone reached AUROC 0.931 within 1 kHz, exceeding every detector there.

Conclusion: On these corpora, measured detector performance depends on how the label is constructed and who provided it, at magnitudes equal to or larger than the differences between detectors. The label acts chiefly through what detectors learn, and the proxy leaks the input. We claim neither label is correct; we show that the choice determines the answer and release the mappings and a CPU-reproducible pipeline.

KEYWORDS

auditory brainstem response, benchmark validity, evaluation methodology, inter-rater reliability, label quality, objective detection

1 Introduction

The auditory brainstem response (ABR) underpins objective audiometry wherever behavioral thresholds are unreliable in newborns, developmentally delayed patients, and in medico-legal assessment. The operational question at each stimulus level is binary: Is a response present? In routine practice, this is settled by visual inspection, a judgement that varies within and between experienced clinicians, particularly near threshold (Vidler and Parker, 2004; Zaitoun et al., 2014). Automating it has been an active goal for four decades, first through statistical detectors such as Fsp, Fmp, and template cross-correlation (Elberling, 1979; Elberling and Don, 1984; Sininger, 1993; Cebulla et al., 2000; Chesnaye et al., 2018), later through automated quality assessment and threshold-seeking procedures (Valderrama et al., 2014; Suthakar and Liberman, 2019), machine learning on engineered features (Acir et al., 2006; Davey et al., 2007), and, most recently, deep networks applied to the waveform (McKearney and MacKinnon, 2019; McKearney et al., 2022; Liu et al., 2025). Across supervised learning more generally, the quality of the target has received far less attention than the choice of model; when it has been examined, the consequences have been large: Label errors in widely used test sets change which model ranks first (Northcutt et al., 2021), benchmark labels have been found to be ambiguous enough that "progress" partly tracks annotator conventions (Beyer et al., 2020), and public medical imaging datasets have been shown to carry labels that do not mean what users of the data sets assume (Oakden-Rayner, 2020). Audiometric machine learning has not been exempt from this pattern (Sadegh-Zadeh et al., 2024).

Progress in objective ABR detection is measured largely on two small public corpora. The PhysioNet earndb set of Silva and Epstein (2010) contributes tone-pip ABR from eight normal-hearing listeners and has been used as training or test data by McKearney and MacKinnon (2019) and Liu et al. (2025); the Mendeley set of Wang et al. (2021) contributes click ABR from eight further listeners. Both are freely available, both are small, and importantly, neither distributes expert annotations as part of the primary release. Table 1 lists the published detector evaluations on these corpora that we could identify, with the source of the label each used. Every one of them relied on clinician annotation that the authors commissioned or produced themselves: McKearney and MacKinnon (2019) classified the paired earndb replicates visually into three classes; Wang et al. (2021) had five audiologists assign thresholds to their own recordings; Liu et al. (2025) commissioned one audiologist for earndb and a sixth for the Mendeley set and were the first to publish the annotations. A reader of the public releases before 2025 therefore had no annotation to use. The one label that can be computed from the released files alone is an automatic proxy, and the natural proxy is the clinical two-replication reproducibility criterion: Two independent sub-averages are recorded at each level, and a response is deemed present when they correlate above a threshold. Cross-correlation between replications is well-established as a response-detection statistic in its own right (Berninger et al., 2014; Wang et al., 2021), which is part of why it is a natural candidate for an automatic label. The proxy is attractive because it is objective, cheap, and rooted in clinical practice.

We do not claim that this substitution is common in the published literature; Table 1 shows that it is not. We claim that it is available, it is the only label the public releases make available, and its behavior as a research target is untested. That behavior is a separate question from the proxy's utility as a clinical heuristic, and testing it has recently become possible. Liu et al. (2025), reporting a multi-center comparison of nine deep models trained on 128,123 clinically annotated waveforms, published as Supplementary material the audiologist annotations they commissioned for both public corpora: one invited rater for earndb and six raters for the Mendeley set. These annotations can be mapped back onto the original recordings, allowing the proxy and expert judgments to be compared directly on identical data, and allowing detector comparisons to be repeated under each label in turn.

We report that comparison, structured around three questions: First, how well does the proxy agree with expert judgement, and what does the proxy actually measure? We find an agreement of $\kappa = 0.17$–0.25 against $\kappa = 0.69$ between experts, and we show, analytically, by simulation and on the recordings themselves, that the proxy is largely a function of signal-to-noise ratio: a single-replicate instrument statistic predicts it well even where the expert sees no response. Second, through which channel does the label change a study's conclusion? A 2×2 factorial over training label and evaluation label shows that the effect runs chiefly through training: Detectors trained on the proxy lose 0.14–0.30 pooled

TABLE 1 Published detector evaluations on the two public corpora, and the source of the label each used.

References	Corpus	Label source	Raters	Unit labeled	Annotation released?
McKearney and MacKinnon (2019)	earndb (232 paired waveforms)	Visual classification by the authors into clear response/inconclusive/absent	1 (authors)	Paired replicates	No
Wang et al. (2021)	Mendeley (own recordings)	Audiologists' visual thresholds, used to validate a cross-correlation threshold method	5	Threshold per ear, implying presence above it	Thresholds reported in the article
Liu et al. (2025)	earndb (208 waveforms); Mendeley (95 waveforms); also Southampton	Commissioned audiologist annotation (external validation of clinically trained models)	1 (earndb); 6 (Mendeley, five from Wang et al. plus one)	Single waveform	Yes (Supplementary Tables S1–S8)
This study	earndb (204 conditions); Mendeley (86 conditions)	(a) Liu et al. annotations mapped to records; (b) two-replication reproducibility surrogate computed from the released files	1 (earndb); 6 (Mendeley); surrogate has none	Condition (both sub-averages)	Mappings released with this article

To our knowledge, these are the only published detector evaluations on either corpus; none used an automatic proxy, and no annotation was distributed with either corpus before Liu et al. (2025) published theirs as Supplementary material.

AUROC against expert judgement, a loss that arises when records are compared across listeners and frequencies rather than within them. Third, how large is the resulting label × detector interaction, and how much of it is an artifact of pooling across acquisition conditions? Pooled over carrier frequency, the interaction is large and reorders the detectors; approximately a third of it is attributable to pooling, and the remainder is of the same order as the smallest difference eight listeners can resolve.

We are explicit about what this does and does not establish. We do not claim that expert visual judgement is a gold standard: It is itself a proxy for physiological response presence, with known intra- and inter-observer variability, and our own rater analysis quantifies that variability. We therefore make no claim that either label is correct. The claim is conditional and, we argue, more useful: The choice of target determines the conclusion, the two available targets disagree profoundly, and a reader given only a ranking cannot tell which target produced it. Our contribution is that measurement, together with reporting practices that would make such comparisons interpretable, and a release of the label mappings and pipeline.

2 Materials and methods

2.1 Corpora

earndb (PhysioNet). The averaged-waveform ABR from eight normal-hearing adults (four female, 19–31 years), recorded with tone-pip stimuli at 1 and 4 kHz presented at 23.97/s from below threshold to 100 dB peak equivalent sound pressure level (peSPL) in 5 dB steps (Silva and Epstein, 2010). Signals were sampled at 48 kHz and band-pass filtered at 30–3,000 Hz; each stored epoch spans 0–41.7 ms after stimulus onset. Each condition is stored as two independent sub-averages, R1 and R2, each with its own header, including the instrument's wnsfmp value computed for that sub-average alone. Of 232 conditions with a complete R1/R2 pair, 204 carry an expert annotation and form the analyzed set; the 28 excluded conditions lie at 10–45 dB, below the range covered by the published annotation table.

Mendeley (Wang et al., 2021). Click ABR from eight adults aged 23–73 years, recorded on an Intelligent Hearing Systems SmartEP at 20 kHz with 100–3,000 Hz band-pass filtering, stimuli at 37.1/s from 60 to 0 dB in 5 dB steps. Each trace stores an average together with two buffers, the interleaved odd- and even-sweep accumulations that the SmartEP software maintains as independent halves of the same acquisition, which serve as the two sub-averages. Of the available conditions, 86 carry annotations from all six raters and form the analyzed set.

Both corpora are public; this secondary analysis required no new data collection or ethical approval.

2.2 Labels and validation of the annotation mapping

Reproducibility surrogate. For each condition, we computed the Pearson correlation between the two sub-averages over the

response window (4–16 ms for earndb; 2–10 ms for the shorter-latency click responses in the Mendeley set) of the recorded, acquisition-band waveform, and labeled the condition response present when the correlation exceeded $c = 0.25$. This is the two-replication criterion of clinical practice expressed as a fixed rule; the cross-correlation between interleaved replicates is an established detection statistic (Berninger et al., 2014; Wang et al., 2021), and we chose it for exactly that reason. Our question is not whether the correlation carries information about response presence, which it does, but how a binary label derived from it behaves as the target of a detector comparison. Sensitivity to c is examined in Section 3.9.

Expert annotations. Liu et al. (2025) commissioned audiologist annotations for both corpora and published them in Digital Content 1 in the Supplementary material. For earndb, Table 2 gives one invited audiologist's present/absent judgement per participant, carrier frequency, and stimulus level. For the Mendeley set, Table 3 gives six independent audiologists' judgments per participant and level; five originate with Wang et al. (2021), and the sixth was contributed by Liu et al. We use the majority vote across the six as the consensus label and additionally analyze each rater separately.

Mapping validation. The published earndb table heads its level column "dB SL," whereas earndb levels are recorded in dB peSPL; a systematic offset between the two scales would invalidate the mapping while still producing a superficially complete match. We therefore validated the mapping in five ways. First, all 204 annotated (participant, frequency, level) triples matched an existing record, with no unmatched annotations. Second, the participant identifiers in the published table (N1–N8) are the identifiers under which the PhysioNet records are filed, and the annotated series start at different levels for different participants (25, 30, 40, or 50 dB), while all end at 100 dB; a per-participant sensation-level scale would start every series at the same value. Third, we tested alternative alignments: shifting the annotation levels by −10, −5, +5, and +10 dB reduced the number of matched records to 184, 193, 186, and 170, respectively, so no offset matches as well as the identity mapping; Section 3.2 extends this grid to −40…+20 dB. Fourth, over the 25–100-dB range covered by the table, the pattern of which (participant, frequency, level) cells are populated agreed between the annotation table and the corpus record inventory in 92.6% of cells (237 of 256). All 19 discordant cells are records present in the corpus but absent from the annotation table; there is no cell in which an annotation exists without a corresponding record, which is the direction that would indicate a broken mapping. Fifth, the table's levels span 25–100 dB in 5-dB steps and reach exactly 100 dB for every participant. The corpus maximum is 100 dB peSPL, and Liu et al. (2025) themselves describe the PhysioNet stimuli in their Methods section as running "to 100 dB peak-equivalent SPL"; a column in sensation level that reached 100 dB for every participant would require stimuli of roughly 125–130 dB peSPL, which were never presented. We therefore treat the column heading as a labeling slip and the values as recorded stimulus levels.

TABLE 2 Design analysis: pairwise detector comparisons on earndb under the expert label, with each pair's own standard error and minimum detectable effect (MDE) at 80% power, and the listeners required to detect a true difference of 0.05 at that pair's standard error.

Comparison	ΔAUROC	95% CI	*SE*	MDE	Listeners for Δ = 0.05	Adjusted *p*
LR-eng–wnsfmp	0.171	[0.106, 0.247]	0.034	0.096	30	0.003
MLP-raw–wnsfmp	0.229	[0.149, 0.286]	0.034	0.097	30	0.003
MLP-raw–GBDT-eng	0.140	[0.065, 0.245]	0.045	0.126	51	0.002
LR-eng–GBDT-eng	0.082	[0.008, 0.164]	0.040	0.111	40	0.093
MLP-raw–LR-eng	0.058	[−0.025, 0.127]	0.040	0.111	40	0.192
GBDT-eng–wnsfmp	0.089	[−0.020, 0.175]	0.048	0.136	59	0.192

Paired subject-clustered bootstrap, Holm-adjusted within this family of six.
Observed differences exceed the pair's own MDE for exactly the three significant comparisons. Confidence intervals are for the absolute difference in the direction shown.

TABLE 3 Detector discrimination (AUROC) under each label, each detector trained and evaluated under the label against which it is scored, by leave-one-subject-out cross-validation.

Corpus	Detector	Surrogate label	Expert label
earndb (204 conditions, 8 listeners, 1 rater)	wnsfmp	0.827 [0.765, 0.893]	0.662 [0.617, 0.728]
	LR-eng	0.752 [0.695, 0.812]	0.837 [0.808, 0.876]
	GBDT-eng	0.697 [0.654, 0.764]	0.757 [0.670, 0.816]
	MLP-raw	0.713 [0.659, 0.762]	0.898 [0.847, 0.945]
Mendeley (86 conditions, 8 listeners, 6 raters)	RMS-SNR	0.752 [0.618, 0.866]	0.681 [0.491, 0.852]
	LR-eng	0.770 [0.652, 0.880]	0.668 [0.476, 0.837]
	GBDT-eng	0.756 [0.664, 0.862]	0.599 [0.431, 0.754]
	MLP-raw	0.730 [0.639, 0.810]	0.567 [0.388, 0.722]

Subject-clustered bootstrap 95% confidence intervals in brackets; all values are read from the released result files and differ from the submitted version in the third decimal of some intervals. On earndb, the leading detector changes, and the two learned detectors, LR-eng and MLP-raw, exchange places. Mendeley point estimates also reorder, but no Mendeley comparison is statistically resolved.

The annotating authors have since confirmed both points directly (personal communication): The unit of Table 2 is dB peSPL, and the participants are the eight participants of the PhysioNet release.

Status of the expert label. The expert annotation is a human judgement about the recorded waveform. It is a conventional supervised target and is not an arithmetic function of the model input. It is not, however, a physiological gold standard, and we do not treat it as one; Section 3.6 quantifies its variability directly.

Status of the surrogate label. The surrogate is different in kind. It is computed from the same two sub-averages that the detectors receive as input, so the target is an arithmetic function of the input. Let $R1 = s + n1$ and $R_2 = s + n_2$ be the two sub-averages, where s is any component common to both replicates and $n1$ and n_2 are independent residual noise of variance σ^2. The expected correlation between R1 and R_2 over a window is approximately $\sigma^2{}_s / (\sigma^2{}_s + \sigma^2) = \mathrm{SNR} / (1 + \mathrm{SNR})$, where $\sigma^2{}_s$ is the variance of the common component and $\mathrm{SNR} = \sigma^2{}_s / \sigma^2$. A single-replicate quality statistic of the Fmp family estimates the same ratio, using the residual-noise variance that the instrument accumulates across sweeps. Two consequences follow and are tested in Section 3.3. A detector given one replicate and asked to predict the surrogate is being asked to predict a monotone function of a quantity that an signal-to-noise ratio (SNR) statistic already estimates, and because s is any common component, not only a brainstem response, stimulus-locked residue or slow shared activity raises both the surrogate and the SNR statistic in the absence of any response. Both sub-averages enter the detectors as separate samples carrying the same derived label, which compounds this, and is why we repeat the main analyses at condition level (one sub-average per condition).

2.3 Detectors

Both sub-averages of a condition enter as separate single-run inputs carrying that condition's label, giving 408 samples on earndb and 172 on the Mendeley set. Therefore, every detector observes one average run and must decide presence from it alone.

From each sub-average we extracted six scale-invariant descriptors over a response window and a noise window (25–41 ms for earndb, 12–22 ms for the Mendeley set): response-to-noise ratios of root-mean-square (RMS) amplitude, peak-to-peak amplitude and maximum absolute amplitude; the dominant peak latency; the fraction of spectral energy in 100–1,500 Hz; and the spectral centroid. Engineered features were computed on a zero-phase fourth-order Butterworth 100–1,500-Hz band-passed waveform; waveform-input detectors received the recorded waveform resampled to 220 points and z-normalized.

Four detectors span the capacity range:

- wnsfmp, the weighted normalized signal-to-noise statistic computed by the recording instrument and stored in the header of each sub-average, a member of the residual-noise family of quality measures, was used to decide when averaging may stop (Elberling and Don, 1984; Don and Elberling, 1994). No training is involved. On the Mendeley set, which has no equivalent field, the response-to-noise RMS ratio, an analog of the F_sp family of single-point variance-ratio statistics (Sininger, 1993), serves the same role and is denoted RMS-SNR. We include it not because it has been proposed as a competitive detector—it has not—but because it is the natural non-learned reference, is available without any modeling, and is the quantity a reproducibility-derived label most closely tracks.
- LR-eng, L2-regularized logistic regression ($C = 1$), was used on the six engineered features, plus a per-fold matched-filter score, the correlation of the response window with a template formed from training-fold response-present waveforms.
- GBDT-eng, histogram gradient boosting (depth 3, 200 iterations, learning rate 0.06), was used on the engineered features.
- MLP-raw, a 128–64 multilayer perceptron ($\alpha = 10^{-3}$, 800 iterations), was used on the 220-point waveform.

All hyperparameters were fixed *a priori* from conventional defaults and were not tuned on these data; no feature selection or architecture search was performed. We fixed them deliberately: Tuning inside eight listeners would either require nested cross-validation that leaves too few clusters per fold to be meaningful or introduce exactly the optimism this paper is about. The consequence is that absolute AUROC values below are those of untuned reference implementations. The paper's claims, however, concern differences between labels holding the detector fixed, and every detector is held identical across the two labels, so no comparison reported here depends on the detectors being optimal.

2.4 Evaluation and statistical analysis

All learned detectors were evaluated by leave-one-subject-out cross-validation so that no listener contributed to both training and test. Random splits over records rather than listeners would place statistically dependent observations on both sides of the partition, a failure mode implicated in a broad reproducibility problem across machine learning–based science (Kapoor and Narayanan, 2023). Feature standardization, the matched-filter template, and any calibration map were fitted within training folds only. wnsfmp and RMS-SNR require no training; for discrimination, they are scored raw because a monotone recalibration cannot change the area under the curve within a fold, whereas per-fold calibration maps are not globally monotone and would distort it.

Factorial design. Each learned detector was trained under each label, and its out-of-fold predictions were scored against each label, giving four cells per detector: trained on the surrogate and scored against the surrogate (S→ S), trained on the surrogate and scored against the expert (S→ E), and E→ S and E→ E. The diagonal cells, S→ S and E→ E, are what a study adopting that label would report, and the comparison between them is the estimand of the submitted version of this article. The off-diagonal cells separate two effects that the diagonal confounds. Holding the training label fixed and switching the evaluation label (S→ S vs. S→ E, or E→ S vs. E→ E) measures how the two labels ranked the same fixed predictions; this is the contrast under which no detector adapts, and in which wnsfmp, which cannot be trained, is no longer anomalous. Holding the evaluation label fixed and switching the training label

(S→ E vs. E→ E) measures what training on the proxy costs compared to expert judgement, a within-detector paired difference.

Scales. Discrimination is reported as AUROC pooled over the two carrier frequencies, as a study would typically report it, and as a frequency-adjusted AUROC, the pair-weighted mean of the within-frequency AUROCs (the covariate-adjusted AUC of Janes and Pepe, 2008, 2009, with carrier frequency as the discrete covariate). The adjusted statistic compares response-present and response-absent records only within the same frequency, so it cannot be moved by differences in prevalence between frequencies; the difference between the pooled and adjusted values of any contrast is the contribution of pooling. Within-frequency AUROCs are reported as well.

The label × detector interaction. For each detector d, we define Δ_d = AUROC(E→ E) − AUROC(S→ S) and test whether Δ_a − Δ_b differs from zero for each detector pair. This directly tests the article's claim: that the effect of changing the label is not common to all detectors, and therefore reorders them, rather than inferring it from a change in rank order. Algebraically, the difference-in-differences ΔΔ = (Δ_a − Δ_b) is identical to the change in the pairwise difference (AUROC_a − AUROC_b) under the expert label minus the same difference under the surrogate label, so the interaction and the rank reversal are one quantity; a sign change of the pairwise difference is a reversal, and ΔΔ is its magnitude. Because AUROC is bounded, equal differences at different baselines are not strictly comparable; we therefore also report the interaction on the logit scale as a sensitivity and, for each pair, report the bootstrap probability that the pairwise ordering differs between labels. Because the difference-in-differences is a within-detector contrast of between-label differences, any component of the label change common to all detectors, including any difference in the intrinsic difficulty or prevalence of the two targets, cancels; this is why we do not compare AUROC values across labels directly. The two labels define different prediction problems with different prevalence (39% vs. 81% response-present on the analyzed earndb conditions), so their AUROCs are not on a common scale, whereas the interaction and induced change of rank order are scale-free.

Inference. With eight listeners, inference is the weakest part of any analysis of these corpora, and we state how it is done. The primary test of every contrast is an exact subject-level sign-flip permutation test: Each listener's own contribution to the contrast is computed from that listener's out-of-fold predictions, and all 2n sign assignments are enumerated. Only listeners whose records contain both classes under both labels on the scale in question contribute, which is six listeners pooled and on the adjusted scale, six at 4 kHz and four at 1 kHz, so the smallest attainable two-sided p-value is 1/64 = 0.016 (1/16 = 0.063 at 1 kHz); we do not report any p-value below that floor. Family-wise control across the six pairwise interactions uses the studentised step-down max-T permutation adjustment of Westfall and Young (1993), which respects the discreteness of the permutation distribution; Holm's procedure (Holm, 1979) applied to exact p-values with a floor of 1/64 cannot reach 0.05 for six tests (its smallest attainable adjusted value is 0.094), and we say so rather than apply it. The omnibus p-value for any interaction on each scale is that of the largest studentised statistic in the same max-T permutation distribution. Subject-clustered bootstrap intervals, in which listeners rather than records are the resampling unit (Field and Welsh, 2007), are reported throughout as descriptive summaries of uncertainty, with 2,000 resamples, and are not used as tests; with eight clusters, their coverage is approximate (Varoquaux, 2018). The two procedures do not estimate the same thing: The per-listener contributions of the exact test are computed within each listener's records, whereas the pooled AUROC and its bootstrap interval also compare records across listeners. Where the two disagree, a contrast is present across listeners but not within them, and we say so where it occurs (Section 3.4). The Holm-adjusted bootstrap p-values of the submitted version are retained in Supplementary Table S1 for continuity. Family-wise adjustment is applied within families defined in advance: Holm for the six pairwise detector comparisons under the expert label (Table 2), whose bootstrap p-values are continuous; max-T for the six pairwise interactions on each scale and the three training-label effects.

Two additional analyses guard the inference. First, because both sub-averages of a condition share a label, we repeat the interaction test at condition level, retaining one sub-average per condition (204 independent conditions), which removes any pseudo-replication. Second, we repeat the interaction omitting each listener in turn to confirm that no single listener drives it.

Probability quality is assessed after a subject-disjoint logistic recalibration of the kind introduced by Platt (1999). Calibration is reported as the Brier score, the recalibration slope and intercept from regressing outcomes on the logit of the predicted probability, and expected calibration error under both equal-width and equal-mass binning at 5, 10, 15, and 20 bins.

Because the multilayer perceptron and gradient boosting involve stochastic initialization, both were rerun across 10 random seeds.

2.5 Design analysis

To establish what difference these corpora can resolve, we computed for each detector pair the standard error of the paired subject-clustered AUROC difference separately. The minimum detectable effect at 80% power and a two-sided 5% significance level is

$$\mathrm{MDE} = (Z0.975 + Z0.80)\cdot \mathrm{SE} \approx 2.80 \cdot \mathrm{SE},$$

where Z0.975 = 1.960 and Z0.80 = 0.842 are the standard normal quantiles for the two-sided test size and for power, respectively. Because the standard error of an AUROC difference scales approximately as $1/\sqrt{N}$ in the number of independent listeners, the listeners required to detect a true difference Δ at the same power is

$$N_\Delta \approx N0 \cdot (\mathrm{MDE0}/\Delta)^2,$$

with N0 = 8 and MDE0 the value observed for that pair. Per-pair reporting matters: Standard errors differ across pairs, so a single summary figure would misclassify individual comparisons, and we report the listeners required as a range over pairs rather than as one number.

2.6 Analysis provenance

This is a secondary analysis of existing public data and was not pre-registered. The analysis was developed iteratively, and after inspecting intermediate output, three choices were made: wnsfmp is scored raw rather than through a per-fold calibration map (the latter is not globally monotone and depressed its AUROC from 0.662 to 0.480), the minimum detectable effect is reported per detector pair rather than as a single median (a single figure misclassified one comparison), and the primary LR-eng includes a matched-filter feature, with the alternative reported in Section 3.9. The factorial, frequency-adjusted, degeneracy, and mapping analyses of Sections 3.2–3.5 were added at revision in response to peer review. They are computed from the same out-of-fold predictions as the original analysis; no detector was retrained, retuned, or respecified for them, and the released pipeline regenerates them from the stored predictions. We state these explicitly, report the affected sensitivities, and regard all inference here as exploratory rather than confirmatory.

2.7 Reproducibility

The analyses were performed in Python 3.11 with NumPy, SciPy, scikit-learn and the wfdb package, on CPU only, with random seed 42 unless stated. Scripts and label mappings are released (see Data and Code Availability). Every number in the tables and figures, and every inferential statistic in the text, is read from the result files written by the released pipeline so that text, tables, and figures cannot diverge; the two development-run sensitivities discussed in Section 3.9 are identified as such.

3 Results

3.1 What these corpora can and cannot resolve

We begin with the result that conditions every other one because it is independent of which label is used. Standard errors of the paired subject-clustered AUROC difference ranged from 0.034 to 0.048 across detector pairs under the expert label, giving per-pair minimum detectable differences of 0.096–0.136 at 80% power (Table 2, Figure 1A). Applying each pair's own threshold reproduces the Holm-corrected outcomes exactly: The three significant comparisons (0.171, 0.229, 0.140) exceed their thresholds (0.096, 0.097, 0.126), and the three non-significant ones (0.089, 0.082, 0.058) fall below theirs (0.136, 0.111, 0.111). A single summary threshold would have misclassified the MLP-raw vs. GBDT-eng comparison, which is why we report per-pair values.

Scaling each pair's own standard error, detecting a true difference of 0.10 AUROC at 80% power would require 7–15 listeners (median over pairs 10), a difference of 0.05 would require 30–59 (median 40), and a difference of 0.03 would require 82–164 (median 110; Figure 1B). Two consequences follow. Non-significant comparisons here are design nulls and must not be read as evidence of equivalence. And the rater-induced spread of 0.065–0.082 documented in Section 3.6 sits below the design's resolution, so a study of this size cannot distinguish a real detector difference from the difference between two annotators.

3.2 The surrogate agrees poorly with experts

On earndb, the surrogate and the expert annotation agreed on 52.5% of the 204 analyzed conditions, giving Cohen's $\kappa = 0.167$ (Cohen, 1960), subject-clustered 95% CI [0.058, 0.294], slight agreement on the Landis–Koch scale (Landis and Koch, 1977; Figure 2A). Both sub-averages of a condition carry the same labels, so agreement is a property of conditions, not of sub-averages; the κ point estimate is identical either way, but the counts below are per condition. Disagreement was strongly asymmetric: the surrogate declared a response present where the expert declared it absent in 92 conditions (45% of the analyzed set), the reverse in only 5 (2%). On the analyzed conditions, the surrogate labeled 81% response-present compared to the expert's 39%.

Disagreement concentrated almost entirely at 1 kHz, where agreement fell to 34% ($\kappa = 0.080$) and the surrogate labeled 82% of conditions response-present compared to the expert's 16%; at 4 kHz, agreement was 71% ($\kappa = 0.321$; surrogate 81%, expert 62%). Figure 2B shows the divergence against stimulus level: The labels track one another reasonably at 4 kHz and separate almost completely at 1 kHz.

The same comparison on the Mendeley set, against the majority vote of six audiologists, gave 70.9% agreement and $\kappa = 0.248$. The reference point is what the audiologists achieved with one another: across the 15 rater pairs, the median κ was 0.691, the range was 0.518–0.951, and 26% of conditions were not unanimous (Figure 2C). In both corpora, the surrogate therefore agrees with expert judgement substantially less than experts agree among themselves.

Because the published earndb annotation only covers 25–100 dB, our agreement figure for that corpus is computed on that range. The 28 excluded conditions lie at 10–45 dB and are labeled response-present by the surrogate in only 18% of cases; at such levels, most conditions are response-absent under any criterion, so the excluded conditions are the ones the two labels would largely have agreed. The figure we report is therefore conservative in the unfavorable direction. We report it as measured on 25–100 dB and do not extrapolate it to the full range.

Mapping and plausibility checks. Three further checks bear on whether the expert annotation was aligned to the right recordings and on whether it is physiologically credible. First, extending the level-offset grid, Supplementary Table S3 gives the number of annotations that could match an existing record for each shift of the annotation levels relative to the recorded levels. For shifts of ± 5 and ± 10 dB, the counts are exact (Section 2.2). For larger shifts, the corpus inventory constrains an upper bound: An annotation

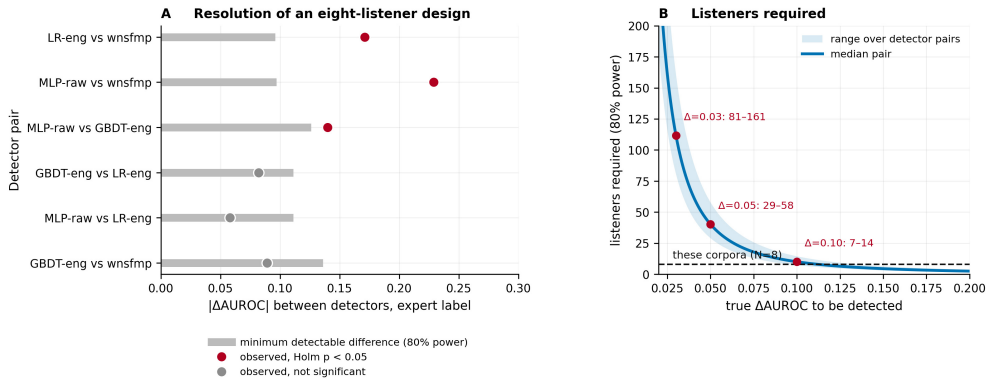


FIGURE 1
What eight listeners can resolve. **(A)** For each detector pair under the expert label, the minimum difference detectable at 80% power (bar) and the observed absolute difference (point; red where Holm-adjusted $p < 0.05$). **(B)** Listeners required to detect a given true AUROC difference at 80% power, for the median detector pair (line) and the range over pairs (band), with the sizes of these corpora marked.

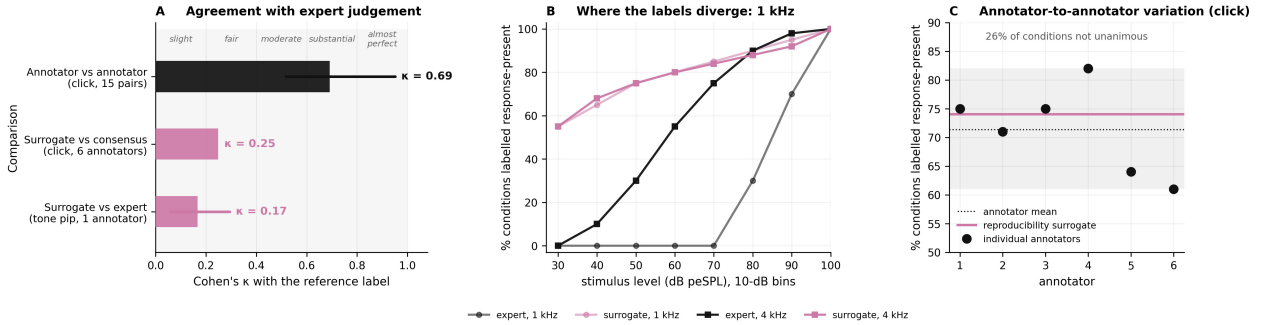


FIGURE 2
The reproducibility surrogate agrees with expert judgement far less than experts agree with each other. **(A)** Cohen's κ between surrogate and expert annotation on earndb (single rater; subject-clustered 95% CI) and on the Mendeley set (six-rater consensus), against the median and range of κ between the six audiologists; shaded bands mark Landis–Koch interpretation ranges. **(B)** Percentage of earndb conditions labeled response-present by label and carrier frequency, in 10-dB bins of stimulus level. **(C)** Percentage of Mendeley conditions labeled response-present by each audiologist (points), their mean (dotted) and the surrogate (line).

shifted to a level above 100 dB peSPL or below the 10 dB corpus floor cannot match, and an annotation shifted to a level between the annotated range and the corpus floor can match only one of the 28 unannotated records, all of which lie at 10–45 dB. Every non-zero shift matches strictly fewer records than the identity (−40 dB: ≤ 110; −35 dB: ≤ 124; −30 dB: ≤ 138; −25 dB: ≤ 152; −20 dB: ≤ 167; −15 dB: ≤ 182; −10 dB: 184; −5 dB: 193; +5 dB: 186; +10 dB: 170; +15 dB: ≤ 154; +20 dB: ≤ 139; identity 204 of 204). Second, for coherence, a per-waveform visual judgement near threshold should occasionally be non-monotone in level, whereas a threshold rule or a misaligned table would not be. Under the expert label 7 of 16 participant × frequency series are non-monotone, in every case by one or two isolated single-level dips (present–absent–present) inside an otherwise monotone series (8 such dips in total; Supplementary Table S2), which is the signature of independent per-waveform judgement rather than of a threshold rule or a shifted table; under the surrogate, 11 of 16 series are non-monotone. Third, physiological plausibility, which deserves direct treatment. The expert annotation records a response at no level in five series (N3 1 kHz, N3 4 kHz, N6 1 kHz, N8 1 kHz, N8 4 kHz); the annotating authors have confirmed that this is indeed their audiologist's judgement for N3 and, as we do, caution that it is a statement about what was visible to that annotator in those recordings, not about the participant's hearing. At 1 kHz, the annotation records responses only at 80 dB peSPL and above in the remaining participants, whereas tone-evoked ABR thresholds in normal-hearing adults are expected some 10–20 dB above behavioral threshold (Stapells, 2000), well below 80 dB peSPL. At 4 kHz, the expert series begin at 30–60 dB peSPL, which is consistent with that expectation and inconsistent with a sensation-level reading of the table, under which nearly every annotated level would lie above threshold. We therefore regard the 4-kHz annotation as physiologically credible and the 1-kHz annotation as conservative relative to physiology. Two readings are possible, and we cannot separate them: The annotator may have applied a strict wave-V morphology criterion that broad, low-amplitude 1-kHz tone-pip responses rarely satisfy at the residual noise levels of these recordings, or the 1-kHz judgments may be unreliable. Either way, the expert label is a visibility judgement on a specific recording, not a statement about physiology, and its reliability cannot be estimated from a single rater; this is the point of Section 3.6. Because the 1-kHz annotation is the questionable half, every result below is

reported within 4 kHz as well as pooled, and none of the article's conclusions depends on 1 kHz alone. Because N3 is the participant with no annotated response at any level, we also recomputed the main quantities with N3 removed (Supplementary Table S7): κ rose from 0.167 to 0.231, the pooled wnsfmp vs. MLP-raw interaction was −0.344 (−0.189 frequency-adjusted); the pooled training-label effects were +0.137, +0.153, and +0.297 for LR-eng, GBDT-eng, and MLP-raw; and wnsfmp predicted the surrogate among expert-absent sub-averages with AUROC 0.777; no conclusion depends on that participant.

3.3 The surrogate largely reflects single-replicate SNR

Section 2.2 derived that the expected correlation between replicates is SNR / (1 + SNR) for any component common to both replicates, response or not. Figure 3A shows the derivation against a simulation of 2,000 conditions in which no response is present at all: Each replicate is a shared non-response component (a slow 30–150-Hz oscillation of condition-varying amplitude, standing in for stimulus-locked residue or shared slow activity) plus independent noise of condition-varying level. The simulation adds nothing to the algebra beyond a check that the finite-window sample correlation and a thresholded label behave as the expectation predicts; the empirical test follows. The simulated correlation follows the analytic curve, and at $c = 0.25$, the surrogate declares a response present in 53% of these response-free conditions. A single-replicate statistic of the Fmp type, the response-window power divided by the residual-noise variance that the instrument estimates across sweeps, predicted the surrogate with AUROC 0.872 with no response present, and 0.940 when responses of varying amplitude were added (Figure 3B); the true common-component SNR, which no single replicate can know exactly, reached 0.901 and 0.952, so the single replicate predicts the surrogate nearly as well as the independent noise of the second replicate allows. A two-window ratio computed from the replicate alone, which normalizes by a late window that also contains the shared component, did not (0.531); the degeneracy is specific to statistics that, like wnsfmp, normalize by a sweep-level noise estimate.

The recordings show the same thing. On earndb, wnsfmp, read from the header of the single sub-average and involving no learning, predicted the surrogate label with AUROC 0.827 over all sub-averages. Restricted to the 250 sub-averages in which the expert saw no response, it still predicted the surrogate with AUROC 0.789 (subject-clustered 95% CI [0.720, 0.860]); the surrogate called 74% of these expert-absent sub-averages present. Within 1-kHz expert-absent sub-averages, the figure was 0.831 ($n = 174$); within 4 kHz, it was 0.689 ($n = 76$). Figure 3C shows why: The 1-kHz sub-averages that the surrogate calls present and the expert calls absent have a mean wnsfmp of 4.13 ($n = 136$, median level 68 dB), higher than the 4-kHz sub-averages in which both labels agree a response is present (3.32). Whatever is reproducible in those 1-kHz recordings, the instrument's SNR statistic and the two-replicate correlation both register it, and the annotator does not call it wave V. This does not require the annotator to be right: If the expert missed genuine low-amplitude responses, the surrogate and wnsfmp agree because both register them; if not, they agree because both register shared non-response structure. In either case, the two quantities are tied to each other rather than to the annotation, which is the point.

This is the structural reason wnsfmp ranks first under the surrogate (Section 3.5): By construction, it estimates the quantity of which the surrogate is a thresholded, noisy realization, and it predicts the surrogate about as well as the noise in the second replicate allows. A rank reversal against the surrogate is therefore not evidence about detectors; it is evidence that the surrogate leaks the input. We regard this, rather than any interaction test, as the article's central finding about the proxy.

3.4 Training on the proxy and scoring against the proxy are different things

Table 4 and Figure 4 give the 2 × 2 factorial. Three patterns are visible. First, the diagonal cells reproduce the study-level comparison: under S→ S wnsfmp leads (0.827) and under E→ E MLP-raw leads (0.898). Second, the off-diagonal cells are low for every learned detector. A perceptron trained on the expert label scores 0.574 against the surrogate and one trained on the surrogate scores 0.594 against the expert, both close to chance; for LR-eng the corresponding values are 0.601 and 0.698, respectively, and for GBDT-eng, they are 0.600 and 0.622, respectively. The two labels define prediction problems that, for the highest-capacity detector, are nearly orthogonal. Third, holding the evaluation label at expert judgement and switching the training label, training on the surrogate cost 0.139 pooled AUROC for LR-eng (bootstrap 95% CI [0.064, 0.224]), 0.135 for GBDT-eng ([0.054, 0.225]), and 0.304 for MLP-raw ([0.236, 0.371]). For MLP-raw, the per-listener contribution was positive in every informative listener (+0.197 to +0.583; exact $p = 0.031$, max-T adjusted over the three learned detectors 0.031; omnibus $p = 0.031$), and for LR-eng, likewise (+0.010 to +0.288; exact $p = 0.031$, adjusted 0.375). Where the cost arises is informative. Within a listener and within a carrier frequency, the two trainings ranked that listener's records about equally well: The per-listener MLP-raw difference within 4 kHz ranged from −0.069 to +0.062 and was positive in only two of six listeners (exact $p = 0.875$), and on the frequency-adjusted scale, no exact test was significant ($p = 0.250$ for each detector), although the estimates comparing records across listeners remained positive (+0.099 [0.022, 0.189], +0.065 [−0.004, 0.135], +0.178 [0.044, 0.288]; within 4 kHz, they were +0.125 [0.026, 0.203] and +0.210 [0.018, 0.375]). The cost of proxy training therefore lies in comparisons across listeners and across frequencies, which is what a single decision threshold applied across patients relies on: A surrogate-trained detector orders a given listener's waveforms by level much as an expert-trained one does, but its score scale shifts between listeners and between frequencies in a way unrelated to the expert's judgement, as a detector that has learned a signal-to-noise quantity (Section 3.3) would be expected to do. A within-listener normalization of scores

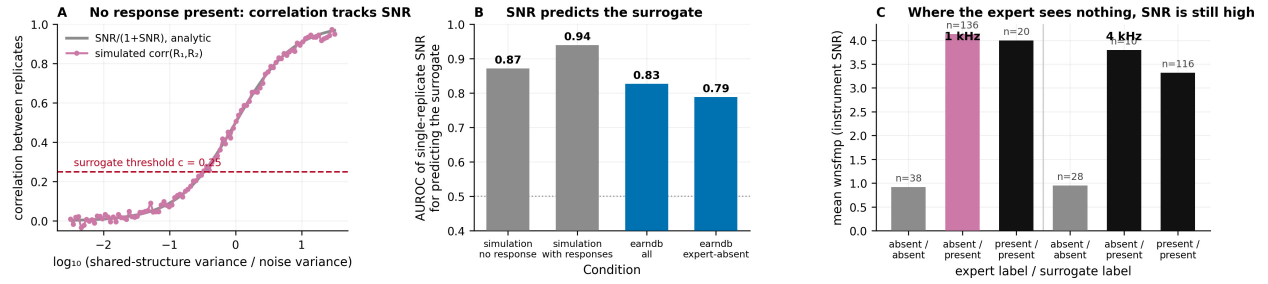


FIGURE 3
The surrogate is largely a function of single-replicate signal-to-noise ratio. **(A)** Simulation with no response present: the correlation between two replicates that share a non-response component follows SNR / (1 + SNR); at the surrogate threshold of 0.25, approximately half of these response-free conditions are labeled response-present. **(B)** AUROC with which a single-replicate SNR statistic predicts the surrogate label in the simulation without and with responses and on earndb over all sub-averages and over the sub-averages in which the expert saw no response (error bar: subject-clustered 95% CI). **(C)** Mean wnsfmp, the instrument's single-replicate SNR statistic, by expert and surrogate label and carrier frequency on earndb; the 1-kHz conditions the surrogate calls present and the expert calls absent have higher SNR than the 4 kHz conditions both labels call present.

TABLE 4 The 2 × 2 factorial on earndb: AUROC of each detector trained under one label (rows) and scored against each label (columns), pooled over carrier frequency and frequency-adjusted (pair-weighted mean of within-frequency AUROCs).

Detector	Trained on	Pooled: vs. surrogate	Pooled: vs. expert	Adjusted: vs. surrogate	Adjusted: vs. expert
wnsfmp	(None)	0.827	0.662	0.824	0.755
LR-eng	Surrogate	0.752	0.698	0.750	0.711
	Expert	0.601	0.837	0.612	0.810
GBDT-eng	Surrogate	0.697	0.622	0.700	0.662
	Expert	0.600	0.757	0.604	0.727
MLP-raw	Surrogate	0.713	0.594	0.713	0.686
	Expert	0.574	0.898	0.575	0.864
Training effect, expert evaluation	LR-eng	+0.139 [0.064, 0.224]; exact $p = 0.031$ (max-T 0.375)		+0.099 [0.022, 0.189]; exact $p = 0.250$ (max-T 0.438)	
	GBDT-eng	+0.135 [0.054, 0.225]; exact $p = 0.062$ (max-T 0.281)		+0.065 [−0.004, 0.135]; exact $p = 0.250$ (max-T 0.531)	
	MLP-raw	+0.304 [0.236, 0.371]; exact $p = 0.031$ (max-T 0.031)		+0.178 [0.044, 0.288]; exact $p = 0.250$ (max-T 0.438)	
	Omnibus (any detector)	exact $p = 0.031$		exact $p = 0.156$	

wnsfmp is not trained, so its rows are identical. The final block gives the training-label effect at fixed expert evaluation, AUROC(E→ E) – AUROC(S→ E), a within-detector paired difference, with subject-clustered bootstrap 95% CI (descriptive, comparing records across listeners), the exact sign-flip permutation p computed from within-listener contributions (six informative listeners, floor 0.016) and its max-T adjustment over the three learned detectors.

would remove the cross-listener component, but it requires a listener's whole level series before any decision is made, which is not how a detector is applied; as produced, the scores are what a study would report. In practical terms, a learned detector trained on the proxy and scored against clinicians reached 0.594–0.698 pooled, within 0.04 of or below the untrained instrument statistic (0.662), and 0.662–0.711 frequency-adjusted, below it (0.755).

The converse contrast, holding the training label fixed and switching only the evaluation label, is the one under which no detector adapts and is the cleanest test of whether the two labels rank the same fixed predictions differently. It did not reorder the detectors at a resolvable magnitude. With training fixed on the surrogate, rescoring against the expert lowered every detector (Δ: wnsfmp −0.165, LR-eng −0.053, GBDT-eng −0.075, MLP-raw −0.119 pooled; −0.069, −0.039, −0.037, −0.027 adjusted), and no pairwise interaction was resolved (wnsfmp vs. MLP-raw $\Delta\Delta = -0.046$ [−0.134, 0.056] pooled, −0.042 [−0.145, 0.118] adjusted; all exact $p \geq 0.094$; Supplementary Table S4). With training fixed on the expert label, the picture is dominated by the near-chance E→ S cells and is uninformative about measurement. The label therefore changes a study's answer chiefly through what the detectors learn, not through how fixed predictions are scored. This is a narrower claim than the one made in the submitted version of this article, and we state it as such.

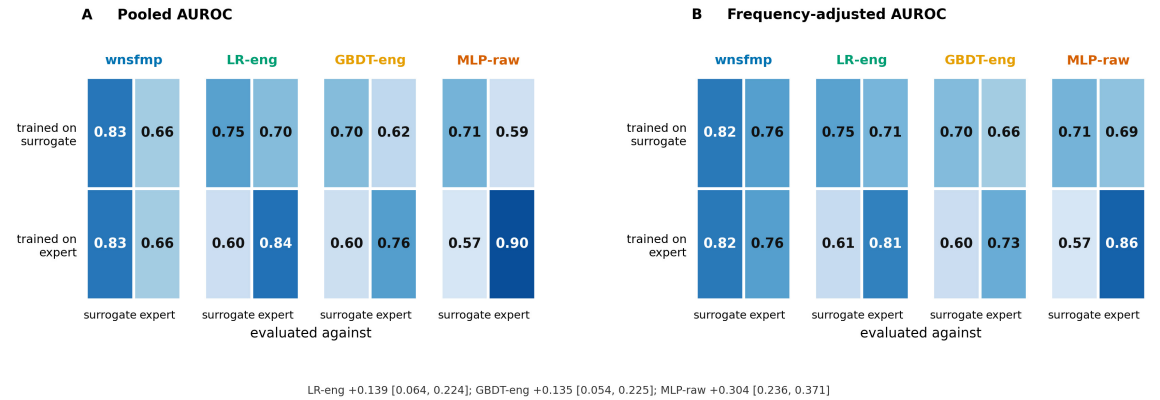


FIGURE 4
Training label and evaluation label have different effects. For each detector, AUROC when trained on the surrogate or on the expert label (rows) and scored against the surrogate or the expert label (columns), **(A)** pooled over carrier frequency and **(B)** frequency-adjusted. Off-diagonal cells are low for every learned detector; the training-label effect at fixed expert evaluation is given beneath the panels with subject-clustered bootstrap 95% intervals.

3.5 The label × detector interaction, decomposed by carrier frequency

Both labels were evaluated on identical data, the same 204 earndb conditions, and the same 86 Mendeley conditions, with each detector trained under the label against which it is scored (Table 3; Figures 5A–C).

On earndb pooled over frequency, the ordering under the surrogate placed wnsfmp first (AUROC 0.827), ahead of LR-eng (0.752), MLP-raw (0.713), and GBDT-eng (0.697). Under expert annotation, the ordering was MLP-raw (0.898, 95% CI [0.847, 0.945]), LR-eng (0.837), GBDT-eng (0.757), and wnsfmp last (0.662). The movement of wnsfmp from first to last is the largest, and Section 3.3 explains it: under the surrogate, wnsfmp is scored against a label that is a thresholded, noisy realization of the quantity wnsfmp itself estimates. The reordering is not confined to it. Among the three learned detectors, which are the systems a comparative study would actually be choosing between, the ranking also changes, with LR-eng leading under the surrogate (0.752 vs. MLP-raw 0.713) and MLP-raw leading under expert annotation (0.898 vs. 0.837).

MLP-raw's pooled AUROC of 0.898 should not be read as a detection result. Carrier frequency alone reaches 0.747 against this label (Section 3.7), and within strata, MLP-raw reaches 0.883 at 1 kHz and 0.853 at 4 kHz, the former being below the stimulus-level baseline. We use the pooled figure only as one endpoint of the label comparison.

Table 5 gives the interaction on each scale, with the exact permutation test as the inferential statistic and bootstrap intervals as description. Pooled over frequency, the wnsfmp-vs.-MLP-raw interaction was $\Delta\Delta = -0.350$ (bootstrap 95% CI [−0.418, −0.268]; exact $p = 0.031$, max-T adjusted $p = 0.031$; omnibus $p = 0.031$), the wnsfmp-vs.-LR-eng interaction, −0.250 (exact $p = 0.031$, adjusted 0.062); and the wnsfmp-vs.-GBDT-eng interaction, −0.225 (exact $p = 0.031$, adjusted 0.094). The interactions between trained detectors were visible in point estimates (LR-eng vs. MLP-raw −0.099, GBDT-eng vs. MLP-raw −0.125) but not resolved by the exact test ($p = 0.094$ for both). The bootstrap probability that the

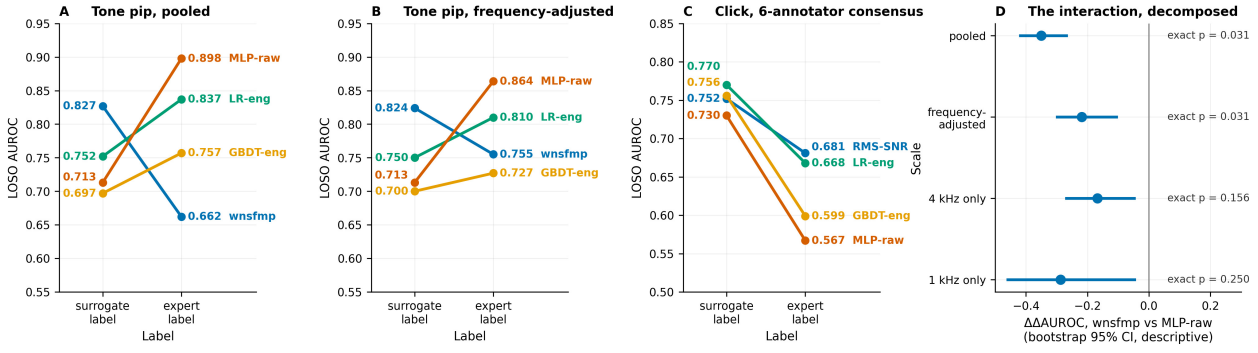


FIGURE 5
Changing the label reorders the detectors, and part of the reordering is pooling. Leave-one-subject-out AUROC under the reproducibility surrogate **(left)** and expert annotation **(right)**, each detector trained under the label against which it is scored, on identical conditions: **(A)** earndb pooled over frequency, **(B)** earndb frequency-adjusted, **(C)** Mendeley against the six-annotator consensus, where point estimates reorder, but no comparison is statistically resolved. **(D)** The wnsfmp vs. MLP-raw interaction pooled, frequency-adjusted and within each frequency, with descriptive bootstrap 95% intervals and the exact permutation *p*-value.

TABLE 5 The label × detector interaction on earndb (each detector trained and scored under the same label), ΔΔ = Δ_a − Δ_b with Δ_d = AUROC(expert) − AUROC(surrogate), on three scales.

Comparison	Pooled ΔΔ [95% CI]	Exact *p*/max-T *p*	Frequency-adjusted ΔΔ [95% CI]	Exact *p*/max-T *p*	4-kHz ΔΔ [95% CI]	Exact *p*/max-T *p*
wnsfmp vs. MLP-raw	−0.350 [−0.418, −0.268]	0.031/0.031	−0.219 [−0.298, −0.103]	0.062/0.094	−0.168 [−0.268, −0.048]	0.156/0.500
wnsfmp vs. LR-eng	−0.250 [−0.344, −0.138]	0.031/0.031	−0.128 [−0.215, −0.009]	0.281/0.688	−0.142 [−0.289, 0.017]	0.344/0.781
wnsfmp vs. GBDT-eng	−0.225 [−0.340, −0.065]	0.031/0.031	−0.097 [−0.200, 0.066]	0.500/0.906	−0.065 [−0.223, 0.128]	0.688/0.906
GBDT-eng vs. MLP-raw	−0.125 [−0.266, −0.029]	0.094/0.156	−0.123 [−0.280, 0.007]	0.344/0.688	−0.103 [−0.219, 0.008]	0.500/0.875
LR-eng vs. MLP-raw	−0.099 [−0.179, −0.022]	0.094/0.094	−0.091 [−0.183, 0.014]	0.594/0.906	−0.026 [−0.135, 0.082]	0.906/0.906
LR-eng vs. GBDT-eng	+0.026 [−0.048, 0.126]	0.875/0.875	+0.032 [−0.037, 0.135]	0.594/0.906	+0.077 [−0.034, 0.191]	0.688/0.906
Omnibus, any interaction	—	0.031	—	0.094	—	0.500

Point estimates are plug-in values; bootstrap 95% intervals (2,000 subject-clustered resamples) are descriptive; the exact subject-level sign-flip permutation p (six informative listeners, floor 0.016) is the inferential statistic, with the single-step max-T family-wise adjustment over the six pairs. The omnibus exact test of any interaction among the four detectors is given in the last row.

pairwise ordering differs between labels was 100% for wnsfmp vs. MLP-raw, 98% for wnsfmp vs. LR-eng, and 81% for LR-eng vs. MLP-raw. On the logit scale, the wnsfmp-vs.-MLP-raw interaction was −2.154 (95% CI [−2.813, −1.620], exact $p = 0.031$).

The pooled interaction is exposed to the pooling artifact described in Section 3.7. Under the expert label, the two carrier frequencies have very different prevalence (16% vs. 62% response-present), and under the surrogate, they do not (82% vs. 81%), so pooling can inflate Δ_wnsfmp under the expert label alone. The frequency-adjusted interaction, which cannot be moved between frequencies by prevalence differences, was −0.219 (95% CI [−0.298, −0.103]) for wnsfmp vs. MLP-raw (exact $p = 0.062$; max-T adjusted $p = 0.094$; omnibus $p = 0.094$), −0.128 (95% CI [−0.215, −0.009]) for wnsfmp vs. LR-eng (exact $p = 0.281$) and −0.097 for wnsfmp vs. GBDT-eng (exact $p = 0.500$). The pooled interaction of −0.350 therefore decomposes into a within-frequency component of −0.219 and a pooling component of −0.130, about 37% of the pooled value. Within 4 kHz alone, where both labels have non-degenerate prevalence, the reordering persisted in point estimates (wnsfmp first under the surrogate, 0.793, and third under the expert, 0.768; MLP-raw last, 0.709, and second, 0.853; LR-eng first under the expert, 0.871) with ΔΔ = −0.168 (95% CI [−0.268, −0.048]), but the exact test did not resolve it ($p = 0.156$). Within 1 kHz, only four listeners have both expert classes, the permutation floor is 0.062, and the interaction (ΔΔ = −0.287, 95% CI [−0.457, −0.050], exact $p = 0.250$) is not estimable in any useful sense (Figure 5D).

The statistically resolved claim is therefore about the comparison a study would actually make, pooled over acquisition conditions. Changing the label reorders the detectors relative to the untrained reference statistic, and a third of that reordering is produced by the prevalence structure the expert label imposes across carrier frequencies, a property of label construction itself. The within-frequency reordering has the same sign and roughly two-thirds of the magnitude, has the same sign in both frequencies (−0.287 and −0.168), is resolved by the exact test for the single contrast most exposed to it but not after family-wise adjustment, and is of the same order as the minimum detectable difference of Section 3.1. We report it as a point-estimate observation with the same status as the trained-vs.-trained interactions.

On the Mendeley set, the point estimates also reorder, and the RMS-SNR reference ranks third under the surrogate (0.752) and first under the six-rater consensus (0.681), while every learned detector falls, but no Mendeley comparison is statistically resolved. With eight listeners and 86 conditions, the subject-clustered intervals overlap extensively (RMS-SNR 0.681, 95% CI [0.491, 0.852]). We therefore report the Mendeley corpus as consistent with label dependence but not as independent confirmation of it, and we draw no inference about the direction of its reordering. Its evidential contribution to this article lies in Section 3.6, where six annotators allow a quantity the earndb corpus cannot supply.

3.6 Rater identity shifts measured performance

The Mendeley annotations allow the label to vary while the detector is held fixed. Rerunning the full leave-one-subject-out evaluation against each audiologist in turn (Figure 6A) moved AUROC by 0.080 for RMS-SNR (0.652–0.733), 0.082 for GBDT-eng (0.591–0.672), 0.066 for LR-eng (0.642–0.708), and 0.065 for MLP-raw (0.561–0.626).

These spreads are of the same order as the quantities the benchmark is used to estimate (Figure 6B). The full spread between the four detectors under the consensus label is 0.113; the mean rater-induced spread within a detector is 0.073, reaching 0.082; and the smallest difference the design can resolve is 0.096–0.136, depending on the pair (Section 3.1). Which clinician annotated the data thus perturbs measured performance nearly as much as replacing one detector with another, and both quantities sit at or below the resolution of the study.

This result also bounds the interpretation of Sections 3.2–3.5, all of which rest on a single earndb annotator. A single-rater target carries the variability of this order, so the earndb comparisons should be read as one draw from an annotator distribution, not as a fixed reference. We return to this in the Limitations section.

3.7 Label-free baselines and pooling artifacts

Two comparators requiring no waveform analysis perform well against the expert label on earndb (Table 6, Figure 7). Stimulus

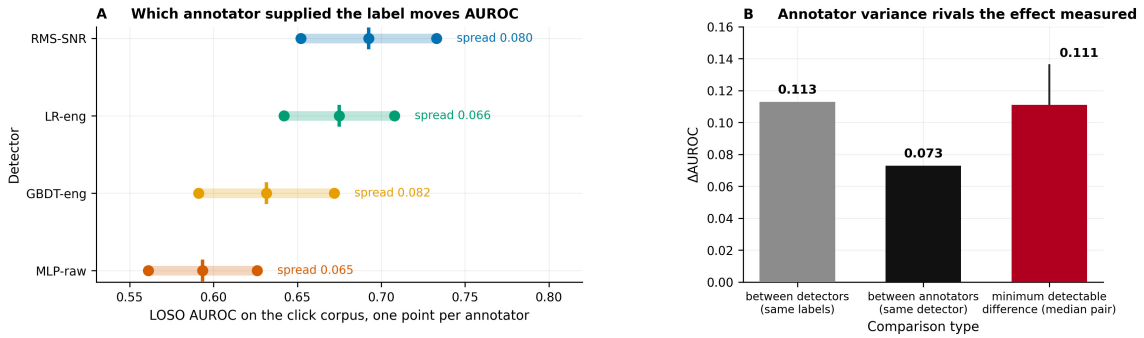


FIGURE 6
Annotator identity moves performance as much as detector choice. **(A)** Leave-one-subject-out AUROC on the Mendeley set with each detector scored against each of the six audiologists individually; each point is one annotator; the vertical mark, the median. **(B)** Spread between detectors under the consensus label, mean annotator-induced spread within a detector, and the minimum difference detectable at eight listeners (median pair; line spans the range over pairs).

TABLE 6 Label-free baselines and detectors against the expert label on earndb, pooled and stratified by carrier frequency.

Comparator	Pooled	Within 1 kHz	Within 4 kHz
Stimulus level alone	0.733	0.931	0.720
Carrier frequency alone	0.747	0.500	0.500
wnsfmp	0.662	0.734	0.768
LR-eng	0.837	0.705	0.871
GBDT-eng	0.757	0.666	0.763
MLP-raw	0.898	0.883	0.853

wnsfmp scores below both strata when pooled (Simpson's paradox). No waveform detector exceeds stimulus level alone within 1 kHz.

level alone reached AUROC 0.733, and carrier frequency alone, predicting response-present for every 4-kHz condition, reached 0.747. Within 1 kHz, stimulus level alone reached 0.931, exceeding every waveform-based detector at that frequency, including MLP-raw (0.883). Only at 4 kHz did learned detectors clearly improve on it (LR-eng 0.871, MLP-raw 0.853 vs. 0.720).

The 1-kHz result admits two readings, and we cannot separate them with these data. Either the expert's 1-kHz judgments were close to a level criterion, which is plausible because 1-kHz tone-pip responses are broad and low in amplitude and clinicians reasonably weight stimulus level when morphology is ambiguous, or 1-kHz response presence is genuinely almost determined by level in normal-hearing ears over this range. On the first reading, the expert label is partly redundant with a stimulus parameter, which would weaken it as an independent target; on the second, the baseline is simply strong. Either way, the practical implication is the same: a detector evaluated on pooled data has not demonstrated waveform-derived information until it is shown to beat the stimulus parameters, which are known to the operator without any analysis.

Frequency-stratified analysis also exposes an artifact of pooling. Against the expert label, wnsfmp achieved 0.734 within 1 kHz and 0.768 within 4 kHz yet 0.662 pooled below either stratum. The mean wnsfmp was 4.13 in the 1-kHz conditions the expert called absent and the surrogate called present, against 3.32 in the 4-kHz conditions both labels called present; pooling therefore concentrates high scores on expert-absent conditions and inverts the relationship, an instance of the reversal described by Simpson (1951). Because detectors are affected to different degrees, pooled comparison across acquisition conditions can reorder them relative to stratified comparison, and Section 3.5 quantifies how much of the pooled interaction this accounts for.

This artifact is not independent of Section 3.2. It arises precisely because the expert label assigns very different prevalence to the two frequencies (16% vs. 62%), whereas the surrogate assigns nearly equal prevalence (82% vs. 81%) and produces no such inversion. The pooling hazard is therefore a consequence of the label's structure rather than a separate phenomenon, and we report it as such. It is a warning about pooling whenever prevalence differs across the pooled conditions, which the surrogate happens to mask.

3.8 Calibration against the expert label

Measured against the expert annotation on earndb (Table 7), the regularized linear model produced the best-behaved probabilities: Brier 0.160, recalibration slope 0.76, intercept −0.08, expected calibration error 0.034–0.073 across binning schemes. MLP-raw achieved the lowest Brier score (0.143) but a recalibration slope of 0.28, indicating the overconfidence characteristic of higher-capacity classifiers (Guo et al., 2017), with an expected calibration error of 0.122–0.140. GBDT-eng was intermediate (Brier 0.191, slope 0.51). Scored as a probability, wnsfmp had the Brier score as 0.249, worse than the prevalence predictor (0.237).

The detector with the strongest discrimination therefore had the least trustworthy probabilities, a dissociation a discrimination-only evaluation would not reveal. These estimates rest on the same eight listeners and the same single annotator as the discrimination results and inherit their uncertainty; they demonstrate the dissociation rather than certifying any detector as clinically calibrated.

At a 0.5 operating point, LR-eng gave sensitivity 0.70, specificity 0.80, positive predictive value 0.69, negative predictive value 0.81, balanced accuracy 0.75, and average precision 0.74. We attempted

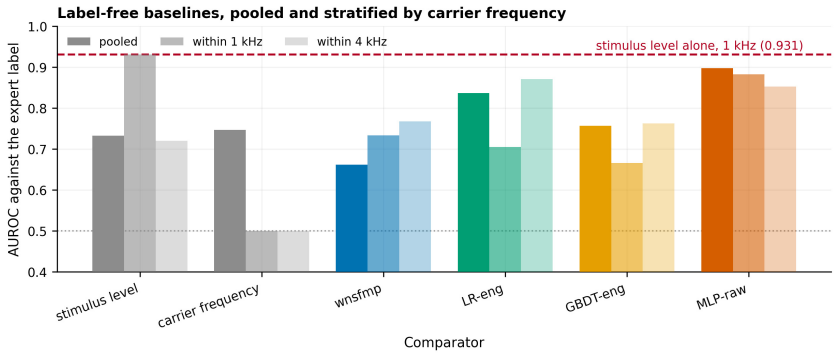


FIGURE 7
Label-free baselines and detectors against the expert label on earndb, pooled, and stratified by carrier frequency; the dashed line marks stimulus level alone within 1 kHz.

TABLE 7 Probability quality against the expert label on earndb.

Detector	Brier	Recalibration slope	Intercept	ECE range
LR-eng	0.160	0.76	−0.08	0.034–0.073
MLP-raw	0.143	0.28	−0.01	0.122–0.140
GBDT-eng	0.191	0.51	−0.25	0.100–0.123
wnsfmp (as probability)	0.249	0.08	−0.43	0.085–0.162
Prevalence predictor	0.237	—	—	—

ECE range spans equal-width and equal-mass binning at 5, 10, 15, and 20 bins.
A recalibration slope below 1 indicates overconfidence. The best-discriminating detector is the most overconfident.

to convert waveform decisions into per-ear threshold estimates, but at 1 kHz, the expert identified no response at any level in several listeners, leaving only 11 ear–frequency combinations; we judged this too few to report and note only that no threshold-level claim is made here.

3.9 Robustness of secondary choices

Across 10 random seeds, MLP-raw gave a median AUROC of 0.894 (range 0.885–0.898, *SD* 0.004) and GBDT-eng was deterministic in practice; the leading result is not a single-seed artifact. Varying the surrogate threshold c over 0.15–0.40 changed its present rate smoothly, without altering any qualitative conclusion, and the surrogate remained far below expert–expert agreement at every value tested.

Removing the paired sub-averages and repeating the pooled interaction test on 204 independent conditions left the wnsfmp vs. MLP-raw interaction essentially unchanged ($\Delta\Delta = -0.321$, subject-clustered bootstrap 95% CI [−0.413, −0.217]) and the trained-vs.-trained interactions unresolved (LR-eng vs. MLP-raw −0.071 [−0.181, 0.045]; GBDT-eng vs. MLP-raw −0.091 [−0.246, 0.052]). Leave-one-listener-out analysis shows every interaction was stable in sign (wnsfmp vs. MLP-raw, −0.364 to −0.338; LR-eng vs. MLP-raw, −0.121 to −0.078; GBDT-eng vs. MLP-raw, −0.170 to −0.082 across the eight deletions), so none is driven by a single listener (Supplementary Table S1). On 204 independent conditions, the leading detector still changes with the label, wnsfmp falling from 0.834 to 0.681 and MLP-raw rising from 0.714 to 0.887.

One analysis choice affects one conclusion. Our primary LR-eng includes the per-fold matched-filter feature (AUROC 0.837); omitting it lowers performance to 0.762, and the LR-eng-vs.-MLP-raw comparison under the expert label, non-significant with the matched filter, becomes significant without it. All wnsfmp-vs.-learned comparisons hold under either variant, as does the interaction test. We report the matched-filter variant throughout and flag this sensitivity explicitly. This sensitivity and the threshold sweep were computed during development with the released code; the pipeline regenerates both (src/ancillary.py) when the corpora are present.

4 Discussion

Objective ABR detection is usually framed as a problem of building better detectors, with progress tracked by accuracy figures on small public corpora. Our results indicate that in those corpora, the measured figures are dominated by decisions about the target rather than by the systems under test.

The reproducibility surrogate is the clearest case. It is a sensible clinical heuristic and an objective, cheap quantity, but as a research target it agrees with expert judgement at $\kappa = 0.17$–0.25 while experts agree with each other at $\kappa = 0.69$. Its failure is structural rather than incidental. The correlation between two replicates is, in expectation, SNR/(1 + SNR) for whatever the replicates

share, and the single-replicate quality statistic printed in the record header estimates the same ratio. On these recordings, that statistic predicts the surrogate with AUROC 0.789 in conditions where the annotator saw no response, and a simulation containing no response at all reproduces the effect. At 1 kHz, tone-pip responses have broad, low-amplitude morphology, and two sub-averages of the same stimulus-locked residue correlate readily; the surrogate consequently declares a response present in 82% of 1-kHz conditions in which the expert, looking for wave V, finds a response in only 16%. A model trained against this surrogate is being trained to predict recording stability, which is related to but distinct from response presence. This reframes the surrogate from "a different question" to a partially degenerate target that leaks the input, and it is why the instrument statistic ranking first under the surrogate is a mechanical expectation rather than an empirical finding about detectors.

The factorial makes the mechanism concrete and, we think, more useful than a rank reversal with a p-value. The same detector trained on the proxy and trained on expert labels, scored against expert judgement, differs by 0.14 (LR-eng), 0.14 (GBDT-eng), and 0.30 (MLP-raw) pooled AUROC for the perceptron in every informative listener. Within a listener and a carrier frequency, the two trainings order waveforms about equally well, so the loss is one of score comparability across listeners and frequencies, which is exactly what a fixed clinical decision threshold depends on; proxy-trained detectors were no better than the untrained instrument statistic against clinicians on the frequency-adjusted scale, and for the perceptron, the two labels defined nearly orthogonal problems (cross-label AUROC 0.594 and 0.574). The converse contrast, fixed predictions scored against each label, did not reorder the detectors at a magnitude this design can resolve. The label therefore acts on a study's conclusion chiefly through what is learned, not through how fixed outputs are scored, and the cost of the substitution is a quantity that can be stated: A detector trained on the proxy loses, against the judgement it is meant to reproduce, more than the differences between competing architectures that these corpora are used to establish.

That distinction determines what a benchmark measures. Under the surrogate, the statistic that best predicts recording stability, the instrument's own signal-to-noise measure, ranks first, and one could conclude that learned detection adds nothing over a number already printed in the record header. Under expert annotation on the same recordings, that statistic ranks last pooled and third in frequency; the waveform network ranks first, and the two learned detectors exchange places. Neither ordering is a computational error; both are correct answers to different questions. The point is that the label sets the question, and a reader given only a ranking cannot tell which was asked.

It could be said that two different targets giving two different rankings is unsurprising. What makes the result worth reporting is by how much, between what, and through which channel. The two targets disagree at $\kappa = 0.167$. Pooled over acquisition conditions, the reordering they produce is large enough to survive an exact subject-level permutation test, and about a third of it is itself a product of label construction. The prevalence structure that the expert label imposes across carrier frequencies and the surrogate does not. The within-frequency remainder has the same sign at both frequencies and is of the order of the smallest difference eight listeners can resolve, which is also the order of the residual noise floor we measure, a rater-induced spread of up to 0.082. The claim is thus quantitative, not conceptual. The substitution that the public releases invite is not innocuous; its cost runs through training, and the corpora cannot absorb the resulting variance.

Our results should not be read as contradicting published performance figures on these corpora. McKearney and MacKinnon (2019) reported 92.9% accuracy on earndb using their own visual annotations, and Liu et al. (2025) reported 94.7% accuracy on the same corpus with a transformer trained on 128,123 clinically annotated waveforms and evaluated against the annotation we also use. Both used expert, rather than proxy, labels, and both are consistent with our finding that expert-labeled earndb is learnable. Our untuned perceptron reaches AUROC 0.898 on the same target. Nor do our results contradict Wang et al. (2021), who showed that thresholds derived from the cross-correlation of interleaved sweeps agree with audiologists' thresholds: a correlation statistic accumulated sweep by sweep and read as a threshold is a different instrument from a fixed-cut binary label on two halves of an average, and the former's clinical value is not in question here. What our analysis adds is that the same corpus scored against a reproducibility proxy would have ranked the same detectors differently, that training on that proxy would have produced detectors that do not reproduce clinical judgement, and that a single annotator's judgement carries variance of the order of the differences being reported.

The rater analysis extends the point to studies that do use expert annotation. Changing which of six audiologists supplied the target moved AUROC by up to 0.082, the same order as many differences reported between competing architectures and below the resolution limit of these corpora. This does not suggest carelessness; median pairwise agreement of $\kappa = 0.691$ (8%7–99% raw agreement) is typical for near-threshold ABR judgement. It means that when the annotation carries this much variance, a single-rater benchmark of this size cannot support a ranking, and reporting against one rater without an uncertainty estimate over raters overstates precision. The numbers from Section 3.1 make the requirement concrete: resolving a true difference of 0.05 AUROC between two detectors at 80% power would need roughly 30–59 listeners, and a difference of 0.03, which is the order of many reported gaps, 82–164; resolving a difference the size of the rater-induced spread would need a corpus several times larger than either of these.

The baseline results point to a reporting gap that is straightforward to close, and the pooling artifact to a second. That stimulus level alone reaches 0.931 within 1 kHz is not a criticism of any detector, but it does mean a pooled figure of, say, 0.85 has not yet demonstrated waveform-derived information. And because acquisition conditions differ in both baseline detectability and prevalence, pooled comparison can reorder detectors relative to stratified comparison, as we observed for wnsfmp; the frequency-adjusted AUROC of Janes and Pepe (2008, 2009) is a ready-made remedy, and we recommend it wherever conditions with different prevalence are pooled.

The difficulty is not specific to ABR. Wherever a clinical target is itself a clinician-assigned judgement or a derived score, the

properties of that target propagate into every performance figure computed against it; label errors and ambiguities in widely used benchmarks have been shown to change model rankings in other domains (Northcutt et al., 2021; Beyer et al., 2020; Oakden-Rayner, 2020). What distinguishes the present case is that the substituted target is not a clinician's judgement at all but an automatic statistic, that the statistic is a function of the model input, and that a published annotation exists against which it can be checked.

Our findings are consistent with the direction the field has taken. Liu et al. (2025) argued that generalizable ABR detection requires large, heterogeneous, expertly annotated data. Our analysis supports that conclusion from the opposite end. We quantify what the small corpora cannot deliver and locate the shortfall in label validity and label variance as much as in sample size.

The contrasts this design leaves unresolved—the within-frequency remainder and the trained-vs.-trained interactions—mark the limit of what eight listeners can support under the present frequentist treatment. A hierarchical Bayesian analysis, with listener- and annotator-level variance components, shrinkage across detectors, and explicit priors, would be the natural framework for extending these comparisons, and we regard it as the principal statistical direction for future work.

We therefore suggest five practices, each derived from a specific result above. First, state the provenance of the target label and, where a proxy is used, report its agreement with expert judgement and whether it is a function of the model input because the proxy examined here agrees at $\kappa = 0.17$–0.25 and is predictable from a single-replicate SNR statistic (Sections 3.2, 3.3). Second, where a proxy is used for training, report the cost of doing so against expert judgement where any expert-labeled subset exists, because that cost was 0.14–0.30 AUROC here (Section 3.4). Third, where multiple annotators exist, report performance against each and treat the spread as part of the uncertainty because that spread reached 0.082 here (Section 3.6). Fourth, report label-free baselines, at minimum stimulus level and, where frequencies are pooled, carrier frequency, because these reached 0.733 and 0.747 pooled and 0.931 within 1 kHz (Section 3.7). Fifth, stratify by acquisition condition or report a covariate-adjusted AUROC alongside the pooled figure because pooling inverted wnsfmp relative to both of its strata and accounted for a third of the headline interaction (Sections 3.5, 3.7). None is expensive; together, they distinguish a ranking that can be interpreted from one that cannot.

5 Limitations

The two corpora contribute eight listeners each, all adults, all recorded on a single device per corpus, with normal hearing in earndb and unreported hearing status in the Mendeley set. Nothing here speaks directly to the infant, pediatric, or pathological populations that motivate objective ABR, and no claim about clinical deployability is intended.

Neither label is a physiological gold standard. Expert visual judgement is itself a proxy with documented intra- and inter-observer variability, which, as shown in Section 3.6, quantifies rather than assumes. Our claims are accordingly comparative: The two targets disagree, and the choice between them changes the ranking and is not a determination that either is correct. A study that could be adjudicated would need an independent physiological criterion, which no public ABR corpus currently provides.

The earndb annotation comes from a single rater, so for that corpus label variance cannot be estimated directly; our estimate comes from the Mendeley set and transfers only as an order of magnitude. The 1-kHz half of that annotation is conservative relative to physiological expectation, and one participant carries no annotated response at any level at either frequency (Section 3.2); the annotating authors confirm that this is their audiologist's judgement on those recordings and not a statement about the participant's hearing; whether it reflects a strict morphology criterion or poor recordings cannot be determined from the published material. We have therefore reported every result within 4 kHz, where the annotation is physiologically credible, as well as pooled, and with N3 removed (Supplementary Table S7). Conversely, the Mendeley corpus, which has six raters, is smaller (86 conditions) and yields no statistically resolved detector comparison. The corpora are complementary in their weaknesses rather than jointly conclusive, and the interaction test rests on earndb alone.

Other investigators commissioned the annotations for a different purpose, and we have no access to the annotation protocol beyond what is published; the earndb table covers 25–100 dB, so our agreement estimate excludes 28 lower-level conditions as quantified in Section 3.2. We validated the level mapping against the corpus record inventory (Sections 2.2 and 3.2): no alternative alignment from −40 to +20 dB matches as many records as the identity, no annotation exists without a corresponding record, the participant identifiers are those of the PhysioNet files, and the table's 100-dB ceiling is the corpus's peSPL ceiling. The published column heading reads "dB SL." while the same authors' Methods section describes the levels as peak-equivalent SPL; at our request, the annotating authors have confirmed that the unit is dB peSPL and that the participants are those of the PhysioNet release, which closes the residual risk in the mapping. The single-rater status of the annotation and the conservatism of its 1-kHz half remain.

All inference rests on eight independent clusters. Subject-clustered bootstrap intervals over eight units have only approximate coverage, which is why we use them descriptively and rest inference on exact subject-level permutation tests, whose own limitations are a floor of 1/64 that leaves family-wise control over six contrasts unattainable, and the fact that the two listeners with no expert-present record (N3, N8) contribute to the pooled estimates but not to the within-listener tests; the design analysis should be read as an order-of-magnitude guide: It assumes standard errors scale inversely with the square root of the number of listeners and ignores the correlation structure between detectors. Our detectors are untuned reference implementations, so absolute AUROC values are not competitive claims; because every detector is held identical across the two labels, the label comparisons are unaffected by this, although we cannot exclude that tuning would change the magnitude of the interaction.

What the interaction analysis does and does not establish should be stated in one place. Pooled over acquisition conditions, the label change reorders the learned detectors relative to the untrained reference statistic, and this is resolved by the exact test. Approximately a third of that pooled interaction is a pooling effect produced by the label's prevalence structure; the within-frequency remainder and every interaction between two trained detectors are visible in point estimates, consistent in sign across listeners and frequencies, and not resolved at this sample size. The contrast in which no detector adapts, fixed predictions scored against each label, is null at this sample size. None of these unresolved contrasts is claimed; they are reported as observations whose resolution would require the listener numbers found in Section 3.1. The resolved findings that do not depend on the interaction test are the label-agreement result, which involves no modeling; the degeneracy of the surrogate, which is analytic and reproduced on the recordings; and the pooled within-detector cost of training on the proxy, with the qualification from Section 3.4 that it arises across listeners and frequencies rather than within them.

The study was not preregistered, and several analysis choices were made after inspecting intermediate results (Section 2.6); the reported inference is exploratory. Finally, we examine one proxy, two-replication reproducibility at a specific threshold and window, and cannot exclude that another automatic proxy would agree better with expert judgement. Establishing that would require the comparison we make here, which is the practice we advocate.

6 Conclusion

Of the two small public corpora most often used to evaluate objective ABR detection, measured performance and detector ranking depend on how the target label is constructed and which clinician provided it. A reproducibility proxy agrees with expert judgement far less than experts agree with one another, and it does so for a structural reason: It is a function of signal-to-noise ratio that a single-replicate instrument statistic predicts well where no clinician-visible response exists. Training detectors on the proxy, rather than expert, labels costs 0.14–0.30 pooled AUROC against expert judgement, a loss of score comparability across listeners and frequencies that leaves learned detectors no better than the instrument statistic; scoring fixed predictions against one label or the other does not reorder them at a resolvable magnitude. Switching labels as a study would, for both training and evaluation, reorder detectors relative to the untrained reference pooled over acquisition conditions, a third of that reordering being the pooling structure the label itself imposes, while the within-frequency reordering and the reordering among trained detectors are visible but unresolvable at this sample size; and changing the annotating clinician moves performance by an amount of the same order as changing the detector. Two label-free baselines are strong enough that at 1 kHz, no waveform detector improves on stimulus level alone. These findings question the reliability of detector rankings derived from these corpora and suggest that such corpora are better suited to methodological study than to model comparisons. We release the label mappings and pipeline and recommend that evaluations report label provenance and proxy validity, the cost of proxy training where it can be measured, rater-level uncertainty, label-free baselines, and stratified or covariate-adjusted rather than pooled comparisons.

Data availability statement

The earndb corpus is available at https://physionet.org/content/earndb/1.0.0/ and the Mendeley corpus at https://data.mendeley.com/datasets/4yb9772dff/1. The expert annotations analysed here were published by Liu et al. (2025) as Supplemental Digital Content 1 to https://doi.org/10.1177/23312165251347773. Machine-readable mappings of those annotations onto the corpus records, the stored out-of-fold predictions, and the scripts that regenerate every figure and reported number are available at https://github.com/salisadegh/AI-in-auditory-neuroscience. A versioned snapshot (Zenodo DOI) will be created on final acceptance and cited in the published version.

Author contributions

S-AS-Z: Project administration, Validation, Data curation, Visualization, Methodology, Formal analysis, Conceptualization, Investigation, Funding acquisition, Writing – review & editing, Supervision, Resources, Writing – original draft, Software. SSa: Visualization, Data curation, Conceptualization, Methodology, Investigation, Validation, Funding acquisition, Supervision, Writing – review & editing, Project administration, Writing – original draft, Formal analysis, Resources, Software. MB: Software, Formal analysis, Writing – original draft, Writing – review & editing, Resources, Data curation, Conceptualization, Methodology, Investigation, Visualization, Supervision, Project administration, Funding acquisition, Validation. SSh: Resources, Funding acquisition, Project administration, Writing – original draft, Visualization, Validation, Formal analysis, Writing – review & editing, Data curation, Supervision, Investigation, Conceptualization, Software, Methodology. DJ: Writing – review & editing, Resources, Software, Formal analysis, Writing – original draft, Visualization, Investigation, Data curation, Validation, Conceptualization, Funding acquisition, Methodology, Supervision, Project administration.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgments

We thank the authors of Liu et al. (2025) for confirming, at our request, the unit and participant correspondence of their published earndb annotation.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fauot.2026.1952229/full#supplementary-material

References

Acir, N., Özdamar, Ö., and Güzeliş, C. (2006). Automatic classification of auditory brainstem responses using SVM-based feature selection algorithm for threshold detection. *Eng. Appl. Artif. Intell.* 19, 209–218. doi: 10.1016/j.engappai.2005.08.004

Berninger, E., Olofsson, Å., and Leijon, A. (2014). Analysis of click-evoked auditory brainstem responses using time domain cross-correlations between interleaved responses. *Ear Hear.* 35, 318–329. doi: 10.1097/01.aud.0000441035.40169.f2

Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and van den Oord, A. (2020). Are we done with ImageNet? *arXiv* [Preprint] *arXiv:2006.07159*. doi: 10.48550/arXiv.2006.07159

Cebulla, M., Stürzebecher, E., and Wernecke, K. D. (2000). Objective detection of auditory brainstem potentials: comparison of statistical tests in the time and frequency domains. *Scand. Audiol.* 29, 44–51. doi: 10.1080/010503900424598

Chesnaye, M. A., Bell, S. L., Harte, J. M., and Simpson, D. M. (2018). Objective measures for detecting the auditory brainstem response: comparisons of specificity, sensitivity and detection time. *Int. J. Audiol.* 57, 468–478. doi: 10.1080/14992027.2018.1447697

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Davey, R., McCullagh, P., Lightbody, G., and McAllister, G. (2007). Auditory brainstem response classification: a hybrid model using time and frequency features. *Artif. Intell. Med.* 40, 1–14. doi: 10.1016/j.artmed.2006.07.001

Don, M., and Elberling, C. (1994). Evaluating residual background noise in human auditory brain-stem responses. *J. Acoust. Soc. Am.* 96, 2746–2757. doi: 10.1121/1.411281

Elberling, C. (1979). Auditory electrophysiology: the use of templates and cross correlation functions in the analysis of brain stem potentials. *Scand. Audiol.* 8, 187–190. doi: 10.3109/01050397909076320

Elberling, C., and Don, M. (1984). Quality estimation of averaged auditory brainstem responses. *Scand. Audiol.* 13, 187–197. doi: 10.3109/01050398409043059

Field, C. A., and Welsh, A. H. (2007). Bootstrapping clustered data. *J. Royal Stat. Soc. B* 69, 369–390. doi: 10.1111/j.1467-9868.2007.00593.x

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning* (Sydney: PMLR), 1321–1330.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.

Janes, H., and Pepe, M. S. (2008). Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. *Am. J. Epidemiol.* 168, 89–97. doi: 10.1093/aje/kwn099

Janes, H., and Pepe, M. S. (2009). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika* 96, 371–382. doi: 10.1093/biomet/asp002

Kapoor, S., and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4:100804. doi: 10.1016/j.patter.2023.100804

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Liu, Y., Xiang, L., Li, Q., Li, K., Yang, Y., Wang, T., et al. (2025). Comparison of deep learning models for objective auditory brainstem response detection: a multicenter validation study. *Trends Hear.* 29:23312165251347773. doi: 10.1177/23312165251347773

McKearney, R. M., Bell, S. L., Chesnaye, M. A., and Simpson, D. M. (2022). Auditory brainstem response detection using machine learning: a comparison with statistical detection methods. *Ear Hear.* 43, 949–960. doi: 10.1097/AUD.0000000000001151

McKearney, R. M., and MacKinnon, R. C. (2019). Objective auditory brainstem response classification using machine learning. *Int. J. Audiol.* 58, 224–230. doi: 10.1080/14992027.2018.1551633

Northcutt, C. G., Athalye, A., and Mueller, J. (2021). "Pervasive label errors in test sets destabilize machine learning benchmarks," in *Proceedings of the 35th Conference on Neural Information Processing Systems, Datasets and Benchmarks Track* (Red Hook, NY: Curran Associates).

Oakden-Rayner, L. (2020). Exploring large-scale public medical image datasets. *Acad. Radiol.* 27, 106–112. doi: 10.1016/j.acra.2019.10.006

Platt, J. C. (1999). "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, eds. A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans (Cambridge, MA: MIT Press), 61–74.

Sadegh-Zadeh, S.-A., Soleimani Mamalo, A., Kavianpour, K., Atashbar, H., Heidari, E., Hajizadeh, R., et al. (2024). Artificial intelligence approaches for tinnitus diagnosis: leveraging high-frequency audiometry data for enhanced clinical predictions. *Front Artif Intell* 7:1381455. doi: 10.3389/frai.2024.1381455

Silva, I., and Epstein, M. (2010). *Evoked Auditory Responses in Normals (earndb)*. Cambridge, MA: PhysioNet.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *J. Royal Stat. Soc. B* 13, 238–241. doi: 10.1111/j.2517-6161.1951.tb00088.x

Sininger, Y. S. (1993). Auditory brain stem response for objective measures of hearing. *Ear Hear.* 14, 23–30. doi: 10.1097/00003446-199302000-00004

Stapells, D. R. (2000). Threshold estimation by the tone-evoked auditory brainstem response: a literature meta-analysis. *J. Speech Lang. Pathol. Audiol.* 24, 74–83.

Suthakar, K., and Liberman, M. C. (2019). A simple algorithm for objective threshold determination of auditory brainstem responses. *Hear. Res.* 381:107782. doi: 10.1016/j.heares.2019.107782

Valderrama, J. T., de la Torre, A., Alvarez, I., Segura, J. C., Thornton, A. R. D., Sainz, M., et al. (2014). Automatic quality assessment and peak identification of auditory brainstem responses with fitted parametric peaks. *Comput. Methods Programs Biomed.* 114, 262–275. doi: 10.1016/j.cmpb.2014.02.015

Varoquaux, G. (2018). Cross-validation failure: small sample sizes lead to large error bars. *NeuroImage* 180, 68–77. doi: 10.1016/j.neuroimage.2017.06.061

Vidler, M., and Parker, D. (2004). Auditory brainstem response threshold estimation: subjective threshold estimation by experienced clinicians in a computer simulation of the clinical test. *Int. J. Audiol.* 43, 417–429. doi: 10.1080/14992020400050053

Wang, H., Li, B., Lu, Y., Han, K., Sheng, H., Zhou, J., et al. (2021). Real-time threshold determination of auditory brainstem responses by cross-correlation analysis. *iScience* 24:103285. doi: 10.1016/j.isci.2021.103285

Westfall, P. H., and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York, NY: Wiley.

Zaitoun, M., Cumming, S., and Purcell, A. (2014). Inter and intra-reader variability in the threshold estimation of auditory brainstem response (ABR) results. *Hear. Bal. Commun.* 12, 59–65. doi: 10.3109/21695717.2016.1110957