A Capacity-Matched Test of Observed EEG Amplitude for Motor-Imagery Error Ranking

Anonymous authors
Paper under double-blind review

Abstract

Selective prediction lets a motor imagery system defer uncertain electroencephalography (EEG) trials. We ask whether an observed-amplitude score helps rank decoder errors beyond class probabilities and available channels. We test this on a task built from participants imagining movements of the left or right hand in the PhysioNet EEG Motor Movement/Imagery Database. We measure the error rate among accepted trials with normalized partial area under the risk-coverage curve (pAURC), from accepting half of all trials to accepting every trial. Lower values mean fewer accepted-trial errors. In a follow-up comparison, a permutation sham had the same number of model coefficients as the amplitude selector. Under synthetic electromyographic (EMG) contamination, the full selector minus sham pAURC difference was 0.021% lower than under clean trials. Its participant-level interval included zero. Refitting decoders on separate source participants also left the result unresolved. The earlier comparison with a smaller selector did not meet its prespecified criterion and changed model size, so it cannot isolate the amplitude score. The evidence does not resolve whether the score improves error ranking.

1 Introduction

An electroencephalography (EEG) decoder sometimes must choose between acting on a prediction and deferring it. Missing electrodes and muscle activity can alter the recorded signal, while a small calibration set leaves limited information about a new participant. Selective prediction ranks trials so that a system accepts a chosen fraction and defers the rest (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2019). A ranking score built from observed signal quality may help, but a comparison with decoder confidence alone cannot establish that it does.

Decoder probabilities reveal class and confidence; the channel mask records available sensors. Adding amplitude and its confidence interaction adds two fitted coefficients.

We study that question on a constructed left-versus-right motor-imagery task from PhysioNet's motor-imagery recordings (PhysioNetMI). The decoder and test trials remain fixed across three logistic selectors fit after decoder predictions were fixed. A matched selector receives decoder probabilities and the exact channel mask. A full selector adds an amplitude-derived score and its interaction with confidence. A sham selector receives a feature group of the same size, formed by conditionally permuting score values. The design pairs clean and synthetic electromyographic (EMG) conditions on participants excluded from selector fitting. The score measures observed amplitude after masking; it does not directly measure electrode contact, artifact identity, or user-facing utility. We also repeat the intact clean and EMG analysis with scores permuted once across all source observations and no correctness labels. This keeps the decoder, targets, seeds, replicates, and feature count fixed while removing the conditional strata from the sham.[1]

Under a fixed decoder and held-out runs, does observed amplitude improve error ranking beyond decoder outputs and channel availability? Figure 1 shows the shared inputs and the sham with the same number of fitted coefficients. We compare paired direct effects on clean and EMG trials and their interaction.

[1]Anonymous source code for the matched-sham comparison and synthetic checks is available at https://anonymous.4open.science/r/iclr27-47c8a1/README.md. The repository does not contain participant-level records or empirical result files, so it does not reproduce the paper's participant-level estimates.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

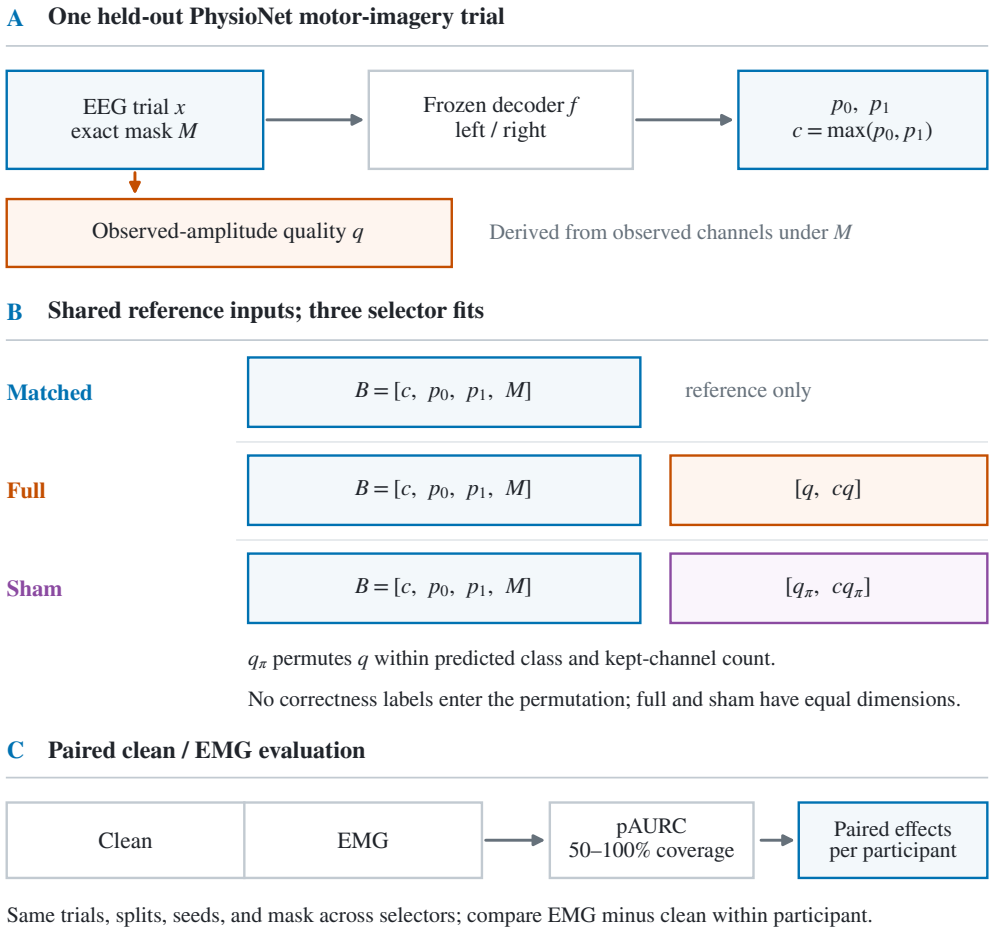


Figure 1: PhysioNet selector comparison protocol. One frozen decoder supplies class probabilities and confidence. The selector also observes the exact binary channel mask for each held-out trial. The matched selector receives these shared inputs. The full selector adds observed-amplitude quality q and its interaction with confidence. The sham with the same number of coefficients replaces q with q_π, permuted within predicted-class and observed-channel-count groups without correctness labels. Clean and synthetic-EMG conditions are paired. The endpoint is normalized partial area under the risk–coverage curve ($\text{pAURC}_{50:100}$) over 50–100% coverage, summarized by paired participant effects. No result is depicted.

Contributions. We make three contributions.

1. A matched selector comparison. It holds decoder probabilities and the exact channel mask fixed while adding one observed-amplitude feature group.
2. Conditional and unconditional permutation shams with the same feature count. This tests the amplitude group against the extra fitted capacity it requires.
3. A paired PhysioNetMI evaluation using held-out runs and selectors fit on source trials. We report the 86 participants with a shared channel layout, an all-89 sensitivity, source-participant deletion, and a refit using a separate source group.

These contributions concern selective ranking under the stated protocol. They do not introduce a new decoder, establish online deferral utility, or resolve how other EEG-quality measures behave.

2 Related Work

Reject-option classification and selective prediction formalize the risk–coverage trade-off (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2019). In motor-imagery EEG, prior work rejects patterns using reliability thresholds or posterior differences, reweights trials during spatial-decoder fitting, or thresholds neural uncertainty (Lotte et al., 2008; Tomida et al., 2015; Ganeshkumar et al., 2017; Milanés-Hermosilla et al., 2021; 2023). These approaches change the reject rule or decoder training. We hold the decoder fixed and ask whether observed amplitude improves held-out error ranking beyond its probabilities and channel mask, using a permutation sham that adds the same two coefficients but replaces their values.

EEG artifacts and preprocessing complicate decoding (Schirrmeister et al., 2017; Uriguen & Garcia-Zapirain, 2015). A recent study evaluates adaptive artifact rejection across 13 motor-imagery datasets (Hajhassani et al., 2026); a recent dual-task study gates commands using 39 differential-entropy features and decoder margin in a subject-specific logistic model (Haridharan et al., 2026). Those methods study epoch rejection, cognitive-load gating, or decoding. We instead test the amplitude score with a sham that adds the same number of coefficients while holding decoder and test trials fixed; other quality measures and quality-aware decoder training remain outside this test.

3 PhysioNet Evaluation Protocol

3.1 Data, participants, and held-out runs

PhysioNetMI comes from the PhysioNet EEG motor movement and imagery collection (Goldberger et al., 2000; Schalk, 2009); MOABB constructs our binary left-versus-right hand-imagery task (Jayaram & Barachant, 2018). The paired study evaluates participants 21–109 (89 targets; 86 share a channel layout), after source-decoder training on participants 1–20. An earlier channel-loss diagnostic instead tests participants 1–20 on held-out run 2 after using runs 0 and 1 for training or adaptation. Its targets overlap the later source-training cohort, but not the later target cohort; its adaptation schedule also differs. We report the effects separately.

For the later paired study, decoder checkpoints are trained on all runs of participants 1–20. Participants 1–15 then supply correctness-labeled trials for selector fitting. Their predictions are therefore in-sample for the source decoder, even though the selector-fitting pass holds out runs during target-style adaptation. Participants 16–20 train the source decoder but do not fit the selectors. Participants 21–109 are absent from source checkpoint normalization and selector fitting. Each target supplies three held-out-run folds, three model seeds, and five paired perturbation replicates. The participant is the inference unit: folds, seeds, replicates, and trials within one participant are aggregated before uncertainty is estimated. Overlapping source training still induces dependence between target fits.

To check whether selector-training labels from the source decoder affect the result, we refit five source decoders per model seed. Each decoder excludes one disjoint triplet among participants 1–15; the excluded triplet then supplies correctness-labeled trials for selector fitting. Participants 16–20 remain in every source training cohort. This makes the selector-fit correctness labels out of sample for their source decoder. The three original target decoder checkpoints, target participants, perturbations, and selector features remain fixed.

A source-metadata audit found that participants 88, 92, and 100 have 128 Hz recordings and 385-sample epochs. The other targets have 160 Hz recordings and 481-sample epochs. The original evaluation included all 89 targets. After the metadata audit, we use the 86 participants with shared channel layout as the primary analysis and retain all 89 as a descriptive sensitivity. The restriction was not specified before the original evaluation. Participant 104 also has fewer recorded trials, so we retain each participant's observed trial count when computing risk.

3.2 Decoder adaptation and paired perturbations

The compact binary decoder is an EEGNet-inspired temporal and depthwise-spatial convolutional network with pooling and a linear classifier (Lawhern et al., 2018). Its mask-aware variant substitutes a learned channel-specific value for unavailable channels. In the initial channel-loss protocol, source training uses 15 epochs and target adaptation uses ten epochs. The later paired study starts from

the existing source checkpoints and adapts each target model for five epochs on eight class-stratified labeled trials from non-held-out runs. Source normalization statistics do not use target evaluation trials. Target correctness labels enter offline scoring only after predictions are fixed; they do not fit the selectors.

The paired study compares clean trials with synthetic electromyographic (EMG) contamination from the public EEGdenoiseNet bank (Zhang et al., 2021). We sample a random 512-sample bank epoch (and one channel if the bank is multichannel), interpolate it linearly over normalized sample positions to the target length, demean it, and divide by its standard deviation plus 10^{-6}. With probability 0.3 per test trial, we add $2s_j w_{ij} r_{ij} e_i(t)$ to channel j, where s_j is the standard deviation of source-normalized test data across trials and time, with 10^{-6} added, $w_{ij} \sim \mathcal{U}[0.25, 1]$, and $r_{ij} \in \{-1, +1\}$ is an independent random sign. Selectors receive identical perturbations and masks; clean trials receive no added waveform. This rescaling changes the bank's physical time scale, so the intervention does not represent natural online EMG. Electrooculographic (EOG) contamination and independent or persistent 50% channel loss are secondary diagnostics. For EOG, the same signed scaling uses weights that decrease linearly from 1.0 to 0.35 over the first four channels; other channels receive no added signal.

3.3 Observed amplitude and selector inputs

Let x_{ijt} be the source-normalized value for trial i, channel j, and time t, and let $M_{ij} \in \{0, 1\}$ indicate an observed channel. With J channels and T time samples, the implemented score is

$$a_i = \frac{\frac{1}{JT} \sum_{j=1}^{J} \sum_{t=1}^{T} |M_{ij} x_{ijt}|}{\frac{1}{J} \sum_{j=1}^{J} M_{ij} + 10^{-6}}, \qquad q_i = \sigma(3 - a_i), \tag{1}$$

where $\sigma(u) = (1 + e^{-u})^{-1}$. Thus a_i is the mean absolute observed magnitude divided by the observed-channel fraction, and higher magnitude produces a lower q_i. The score does not use target-calibration amplitudes and does not identify artifact type or electrode contact.

The decoder supplies class probabilities (p_{i0}, p_{i1}) and confidence $c_i = \max_k p_{ik}$. Write the shared decoder and channel features as $\mathbf{u}_i = [c_i, p_{i0}, p_{i1}, M_{i1}, \ldots, M_{iJ}]$. The matched selector receives $\mathbf{u}_i$. The full selector receives $[\mathbf{u}_i, q_i, c_i q_i]$, and the sham with the same fitted capacity receives $[\mathbf{u}_i, q_{\pi,i}, c_i q_{\pi,i}]$. Here $q_{\pi,i}$ is the score taken from a trial with the same predicted class and number of observed channels. The source permutation uses a fixed seed and stable trial identifiers and never uses correctness labels; a stratum with one trial maps to itself. On targets, a new permutation is drawn within each held-out run, seed, replicate, artifact, and channel-loss batch. A target sham score thus depends on the other trials in its batch; this sham is an evaluation control, not a trial-wise deployment rule. The full and sham selectors have the same feature count and the same non-quality inputs. The control preserves the score distribution within each stratum, but it cannot remove every correlation between quality and correctness.

All learned selectors use the same standardized logistic-regression procedure to predict whether the fixed decoder is correct. They use balanced class weights and identical solver, regularization, and stopping settings. We fit the selectors on the same source-participant observations and score the same held-out target trials. Confidence-only ranking is a descriptive reference: it changes several inputs and fitting steps at once, so its difference from the full selector cannot isolate the amplitude score. The sham with the same fitted capacity was added after the earlier analysis and is a follow-up control, not the originally specified endpoint. The participant bootstrap conditions on the fitted selectors and one permutation per source or target batch; it does not measure variation across independently constructed shams.

Unconditional sham sensitivity. The conditional sham permutes q within groups defined by predicted class and number of observed channels. For a second control, we permute scores once across all source observations, using a fixed seed and no correctness labels. This sham still adds two coefficients, but it does not preserve the conditional groups. The unconditional result is a sensitivity to the sham construction, not a replacement for the originally specified endpoint.

Dependence on selector-training participants. To measure uncertainty from the shared selector-fitting cohort, we omit each of the 15 source participants in turn. We refit all three unconditional

selectors and rescore all 89 held-out targets under the intact clean and EMG conditions. For each target-bootstrap draw b, we add an independent draw $\epsilon_b \sim \mathcal{N}(0, \widehat{\mathrm{SE}}^2_{\mathrm{source}})$, where $\widehat{\mathrm{SE}}_{\mathrm{source}}$ is the leave-one-source jackknife standard error. The combined target-and-source sensitivity interval is the 2.5–97.5% quantile of these sums. It treats the target-resampling and source-jackknife components as independent for this sensitivity construction; it is not a joint repeated-sampling calibration guarantee. This sensitivity leaves the frozen decoder, artifact bank, and recording source shared, so it does not establish generalization to an independent acquisition source.

3.4 Selective-risk endpoint and uncertainty

At coverage c, selector s accepts the $\lceil cn \rceil$ highest-scored trials in one evaluation unit. If r_i indicates whether the decoder is correct, the accepted-trial error risk is

$$R_s(c) = \frac{1}{\lceil cn \rceil} \sum_{i \in A_s(c)} (1 - r_i). \tag{2}$$

The original endpoint is normalized partial area under the risk–coverage curve over $C = \{0.50, 0.60, 0.70, 0.80, 0.90, 1.00\}$:

$$\mathrm{pAURC}_{50:100}(s) = \frac{1}{0.50} \mathrm{Trapz}_{c \in C} R_s(c). \tag{3}$$

Lower values are better. The full and sham selectors have the same number of fitted coefficients. For each participant and condition, we subtract the sham $\mathrm{pAURC}_{50:100}$ from the full selector's $\mathrm{pAURC}_{50:100}$; negative values favor the full selector. The EMG interaction is the EMG paired difference minus the clean paired difference. The originally designated comparison with the smaller selector changes both the inputs and the number of fitted coefficients. It therefore compares feature groups and does not isolate the amplitude score.

We first aggregate seeds, folds, and perturbation replicates within each participant. We then compute paired differences and resample participants jointly across selectors. The bootstrap uses 10,000 draws; a paired Student-t interval is a sensitivity summary. Some target models share source checkpoints and selector-training participants, so the participant intervals do not capture all dependence across targets. Participant-bootstrap intervals hold the fitted selectors fixed. The source-participant jackknife below is a separate refit sensitivity. Intervals for sham comparisons also hold the saved source- and target-side permutations fixed; they do not measure variation from drawing new permutations. The original success criterion required the upper 97.5% bootstrap bound for the EMG interaction to be below zero. Results for participants with a shared channel layout and for the sham with two added coefficients are follow-up analyses; they were not part of the original comparison. Full area under the risk–coverage curve (AURC), area under the generalized risk–coverage curve (AUGRC), and risk at 80%, 90%, and 100% coverage are secondary measures in the protocol. They do not replace the primary partial-risk contrast.

Bootstrap seeds differ by result family. Seed 20260918 was used for the originally designated full-versus-smaller comparison and the unconditional-sham estimates (Table 3). Seed 20260921 was used for the conditional-sham expanded-grid estimates, including the 86-participant primary analysis and all-89 sensitivity (Tables 1 and 2), the coverage analysis, and the target- bootstrap component of the source-participant jackknife. Seed 20260923 was used for intervals after the separate-source refit (Tables 5 and 6). Each bootstrap seed governs participant-level resampling of the corresponding saved effects; the bootstrap does not refit selectors or redraw sham permutations. The jackknife separately refits selectors after omitting each source participant, and the cross-fit analysis uses separately trained source decoders. Sham intervals condition on the saved source- and target-side permutations and do not estimate variation across new draws. Thus, seed 20260918 remains attached to the originally designated and unconditional-sham results; it was not globally superseded by seed 20260921.

Secondary transfer analysis on Cho2017. For each of the 52 Cho2017 participants in turn, a balanced logistic model predicts decoder correctness from 13 standardized trial descriptors and correctness labels from the other 51 participants (Cho et al., 2017; Jayaram & Barachant, 2018). A label-stratified split with seed 0 reserves half of each target's trials for calibration and half for testing. Calibration sets contain 8, 32, or 64 trials; decoder seeds are 0–2. Standard and mask-aware decoders use 15 pretraining and 10 adaptation epochs. Training channel dropout is 0.05 with eight calibration

trials and 0.25 otherwise; test dropout is 50%. We add synthetic EMG from EEGdenoiseNet to 30% of test trials. The descriptors are class probabilities, decoder confidence, amplitude-based quality scores, their confidence interactions, and channel-mask summaries. The target's correctness labels are not used to fit its selector. Appendix results compare ranking by predicted correctness and two averages of percentile ranks, one with confidence and one with quality, against confidence-only ranking at 80% and 90% coverage. These intervals are descriptive and are not adjusted across settings.

4 Results

4.1 Primary comparison from the stored analysis records

The completed paired evaluation contains all 89 participants and three selectors. It covers three artifact conditions (clean, synthetic EMG, and synthetic electrooculographic contamination), three channel states (intact, independent 50% loss, and persistent 50% loss), three decoder seeds, three held-out-run folds, and five paired perturbation replicates.

The primary analysis uses 86 participants with the same channel layout. Participants 88, 92, and 100 have different sampling and epoch lengths, so results for all 89 participants are reported as a descriptive sensitivity. All target predictions, selector-training records, and the aggregate tables use the same saved source-decoder outputs.

The originally designated contrast compared the amplitude selector with a smaller selector that had two fewer fitted coefficients. Across all 89 participants, the change from clean to synthetic EMG in the pAURC difference between these selectors was -0.020% (10,000-draw participant-bootstrap 95% interval $[-0.129, +0.095]\%$). The prespecified rule required the upper 97.5% bootstrap bound to be below zero; this result did not meet it. Because the two selectors differ in both their inputs and coefficient count, this comparison cannot isolate the amplitude feature. A later sham with the same number of fitted coefficients addresses that limitation descriptively.

Table 1 reports the primary analysis. The full selector had a pAURC of 40.259% on clean trials and 42.089% after synthetic EMG, compared with 40.317% and 42.168% for the sham. The full-minus-sham pAURC difference decreased by 0.021% from clean to EMG trials (95% paired participant-bootstrap interval $[-0.129\%, +0.096\%]$). The interval includes zero and does not establish practical equivalence; no minimum effect of interest was specified before evaluation. The sham with the same number of fitted coefficients is a follow-up control, so this is a descriptive comparison. Table 2 shows that retaining the original 89 targets gives a similar, unresolved interaction. Seed provenance is specified in the uncertainty section.

Table 1: Primary analysis after the recording-metadata audit: 86 participants with the same channel layout. Full and sham have the same number of fitted coefficients. Values are normalized pAURC percentages; lower values are better. Differences between the full and sham pAURC values and 10,000-draw paired bootstrap intervals use participant-level values before rounding and condition on the saved sham draws.

Condition	**Full (%)↓**	**Sham (%)↓**	**Difference (%)↓**	**Interval (%)**
Clean	40.259	40.317	-0.058	$[-0.136, +0.018]$
Synthetic EMG	42.089	42.168	-0.079	$[-0.215, +0.050]$
EMG − clean interaction	–	–	-0.021	$[-0.129, +0.096]$

Table 2: Sensitivity retaining all 89 targets from the original evaluation. The sham with the same number of added coefficients is a follow-up control. Entries are differences between the full and sham pAURC values for intact recordings; negative values favor the full selector. Differences and 10,000-draw paired bootstrap intervals use participant-level values before rounding.

Condition	**Full − sham (%)↓**	**95% interval (%)**
Clean	-0.051	$[-0.125, +0.022]$
Synthetic EMG	-0.074	$[-0.204, +0.054]$
EMG − clean interaction	-0.024	$[-0.127, +0.094]$

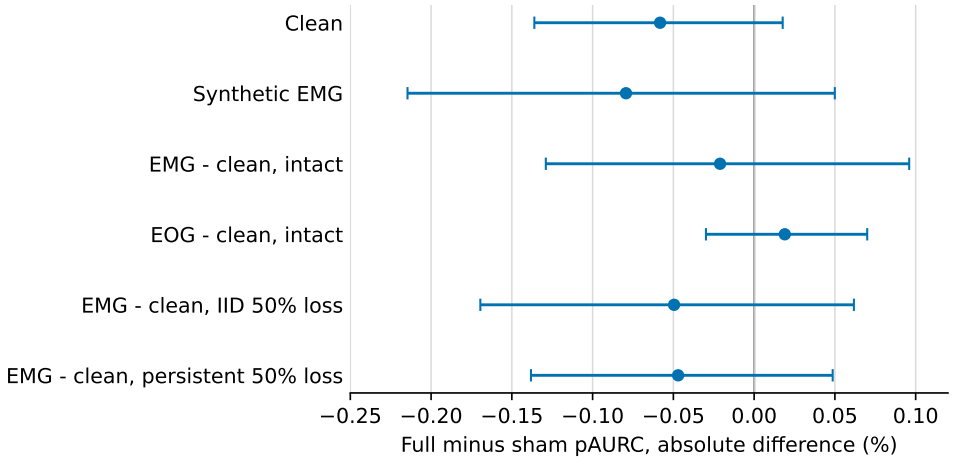


Figure 2: Differences between full-selector and sham pAURC; both selectors have the same number of fitted coefficients for the 86 participants with a shared channel layout. Points are paired participant means; bars are 10,000-draw participant-bootstrap 95% intervals in %. Negative values favor the full selector. The first two entries are direct effects on clean and EMG trials with intact channels. The remaining entries show how the difference changes between clean and artifact conditions. IID denotes independent 50% channel dropout.

Table 5 tests whether source-decoder errors from selector-training participants affect the result. We refit each source decoder without the three participants used to fit that selector, then evaluated the same targets. For the 86 participants with a shared channel layout, the change from clean to synthetic EMG in the full-selector minus sham pAURC difference was -0.043% (95% participant-bootstrap interval $[-0.247\%, +0.159\%]$). Using all 89 participants gave -0.054% ($[-0.258\%, +0.145\%]$). Both intervals include zero and are wider than the original interval. This refit does not resolve an amplitude-score benefit or explain the interval change. The six corresponding EMG or EOG changes in the full-selector minus sham pAURC difference across intact and channel-loss settings also have intervals spanning zero.

The exploratory direct comparisons in Table 6 show lower pAURC for the full selector than for the sham on clean, EMG, and EOG trials under independent 50% channel loss; the nominal intervals exclude zero. All six corresponding intervals using the original source decoders and the three refit intervals under persistent channel loss include zero. We inspected these settings after the analysis and did not adjust for multiple comparisons. They do not show that source refitting caused a gain or resolve whether the effect changes between clean and contaminated trials.

Figure 2 places the two direct effects on intact channels beside the primary and channel-loss interactions. The direct EMG effect has the widest interval, and none of the displayed intervals resolves a pAURC benefit.

4.2 Unconditional-permutation sensitivity

The unconditional sham permutes replacement scores across source trials without preserving the predicted-class and channel-count groups. Table 3 reports the direct effects and changes from clean under synthetic EMG for both target populations. Across all 89 targets, the EMG-related change in the full-selector minus sham pAURC difference was -0.026% (95% bootstrap interval $[-0.137, +0.091]\%$). For the 86 participants with a shared channel layout, it was -0.025% ($[-0.136, +0.096]\%$). Every direct-effect and interaction interval includes zero. This follow-up sensitivity gives the same unresolved conclusion when the sham does not preserve those two groups.

Table 3: Unconditional-permutation sensitivity on intact recordings. Source scores are permuted without using correctness labels or preserving predicted-class and channel-count groups. Entries are full-selector minus sham normalized pAURC differences in %, with 10,000-draw participant-bootstrap 95% intervals computed from unrounded participant values. Negative values favor the full selector. This follow-up control is descriptive.

Population and condition	Effect (%)↓	95% interval (%)
All targets, $n = 89$		
Clean	-0.049	$[-0.118, +0.020]$
Synthetic EMG	-0.076	$[-0.211, +0.060]$
EMG − clean interaction	-0.026	$[-0.137, +0.091]$
Same channel layout, $n = 86$		
Clean	-0.057	$[-0.128, +0.011]$
Synthetic EMG	-0.083	$[-0.220, +0.058]$
EMG − clean interaction	-0.025	$[-0.136, +0.096]$

Table 4 examines how the unconditional-sham result changes when each participant in the selector-training group is left out in turn. The combined intervals, which add a source-participant jackknife term to the target bootstrap, include zero and are wider than the target-only intervals. All 15 deletion estimates for the 89-target EMG interaction are negative, but their range does not establish a stable magnitude. This sensitivity reuses the same source participants and acquisition source; it is not independent acquisition validation or a calibrated joint confidence interval.

Table 4: Sensitivity to leaving out each selector-training participant for the unconditional-sham EMG interaction. Effects, intervals, source standard errors, and ranges over the 15 omissions are %. To form the combined 95% band, each of 10,000 target-bootstrap draws receives an independent normal draw with the source jackknife standard error. The band uses the 2.5–97.5% quantiles and assumes independent components; it is not a joint-calibration statement.

Population	Effect	Source SE	15-source range	Combined 95% band
All targets, $n = 89$	-0.026	0.089	$[-0.112, -0.024]$	$[-0.232, +0.184]$
Compatible, $n = 86$	-0.025	0.102	$[-0.110, -0.011]$	$[-0.260, +0.205]$

4.3 Stress tests and secondary metrics

Appendix Table 7 reports how the full-selector minus sham pAURC difference changes from clean trials to synthetic EMG or EOG, with intact channels and under two forms of 50% channel loss. Every 95% interval includes zero. The aggregate also reports full area under the risk–coverage curve (AURC), area under the generalized risk–coverage curve (AUGRC), and risk at 80%, 90%, and 100% coverage. These secondary measures are descriptive; we did not adjust for the multiple condition and metric comparisons. In an exploratory analysis by coverage, intervals fall below zero at 10–40% coverage but include zero from 50% onward. Table 8 shows the full profile. It does not replace the originally specified pAURC contrast over 50–100% coverage.

The analysis retaining all 89 participants also leaves unresolved how the difference between the full selector and sham changes under synthetic EMG. Its estimate and uncertainty are similar to those for the 86-participant primary analysis. The originally designated comparison with a smaller selector changes both the feature set and fitted capacity, so it does not isolate the amplitude score.

4.4 Initial channel-loss cohort

An earlier diagnostic evaluated 20 PhysioNetMI participants after independent 50% test-channel loss, using eight labeled adaptation trials. It did not include a sham with the same number of fitted coefficients. The difference between full-selector and matched-selector pAURC was $+0.30\%$ (paired participant-level 95% Student-t interval $[-0.18, +0.77]\%$), which leaves the feature-group comparison unresolved (Appendix Table 9).

5 Discussion

The primary 86-participant comparison remains unresolved: direct effects on intact channels and the clean-to-EMG interaction have participant-bootstrap intervals spanning zero. The all-89 interaction and separate-source refit are also unresolved. The same-coefficient sham was added after the original evaluation, so these are follow-up analyses.

Exploratory coverage intervals favor the full selector at 10–40% acceptance but include zero from 50% onward. The originally specified pAURC endpoint covers 50–100% acceptance and remains unresolved. Other risk measures are descriptive and unadjusted across conditions and metrics.

Cho2017 is a secondary selective-accuracy analysis under a separate protocol, with 52 targets in the transfer grid and 20 in an earlier quality-only check. The aggregate records do not establish participant-ID overlap between these analyses, and their selective-accuracy differences cannot be pooled with the PhysioNetMI risk-area differences. Predicted-correctness rank and its equal-weight blend with the stored quality-score rank have positive participant-level 95% interval lower bounds in all 12 settings; the confidence-rank blend does so in 1 of 12. The same 52 participants recur across all 36 settings, without multiplicity adjustment. All 36 settings have files for all 52 targets. The recorded zero-fill sensitivity gives identical estimates and intervals because no target file is missing. Records do not establish that the stored quality field is the primary calibration-referenced amplitude score, and the blend is not compared directly with correctness ranking. An earlier 20-participant quality-only analysis has no positive interval lower bound in four settings; two intervals lie wholly below zero (Appendix Tables 10, 11, and 12).

The disjoint-source refit remains unresolved; sham intervals omit variation across new draws.

Held-out runs and eight labeled adaptation trials do not test action cost, online deferral, or interface utility. One-bank synthetic EMG does not establish natural-artifact behavior. The amplitude proxy does not identify artifact type, impedance, or persistent sensor failure. Intervals omit shared-checkpoint dependence and new sham draws; no independent acquisition source was tested. The channel-loss tests compare intact recordings with two patterns of 50% loss. They do not assess intermediate channel counts or whether the difference between the full selector and sham varies across decoder-confidence strata.

6 Conclusion

The original 89-participant amplitude-versus-smaller-selector comparison missed its prespecified rule: the upper 97.5% bootstrap bound was not below zero. The follow-up same-coefficient sham for 86 shared-layout participants and the all-89 and separate-source interactions remain unresolved. Sham intervals condition on saved permutations. These results do not establish equivalence, performance on natural artifacts, or user-facing utility.

The observed-amplitude score is available from each masked held-out trial and can be examined alongside decoder confidence and channel availability. The present comparisons, however, do not establish that using it improves accepted-trial risk ranking over the specified 50–100% coverage range. A diagnostic score is therefore distinct from a deployed deferral rule: the latter also requires an action cost, a calibrated decision policy, and prospective validation on inputs representative of the intended use.

The matched-sham protocol gives a concrete requirement for that validation. The full selector and its sham should share decoder predictions, target participants, perturbations, fitting conditions, and coefficient count, so that the observed amplitude values are the relevant difference. Multiple independently drawn shams would quantify variation from control construction. Selector fitting on correctness labels from source participants excluded from their decoder's training would further reduce dependence from source-decoder fitting. These controls matter because the exploratory pattern is concentrated at 10–40% acceptance, whereas the prespecified pAURC endpoint averages risk from 50–100% coverage and remains unresolved.

Other observed quality measures, natural artifact sources, and online decision objectives could lead to a different conclusion. Answering those questions requires a prospective study with independent acquisition data and an explicitly defined user cost.

7 Reproducibility Statement

Methods and the appendix specify the cohorts, splits, perturbations, metrics, and analyses. The anonymous code supplement provides the matched-sham implementation and synthetic checks. Raw EEG and participant-level result files are not redistributed; the supplement cannot replay the reported estimates.

8 AI Use Statement

We used generative AI tools to aid writing and editing and for literature retrieval and discovery. The authors reviewed the AI-assisted text and take responsibility for the final paper.

References

Hohyun Cho, Minkyu Ahn, Sangtae Ahn, Moonyoung Kwon, and Sung Chan Jun. Eeg datasets for motor imagery brain–computer interface. *GigaScience*, 6(7):gix034, 2017. doi: 10.1093/gigascience/gix034. URL https://doi.org/10.5524/100295.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11: 1605–1641, 2010. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html.

M. Ganeshkumar, Kai Keng Ang, and Rosa Q. So. Reject option to reduce false prediction rates for EEG-motor imagery based BCI. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2964–2967. IEEE, 2017. doi: 10.1109/EMBC.2017.8037479.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Davoud Hajhassani, Bruno Aristimunha, Paul-Adrien Graignic, Apolline Mellot, Lionel Kusch, Arnaud Delorme, Thomas Semah, and Arnault H. Caillet. From EEG cleaning to decoding: The role of artifact rejection in MI-based BCIs, 2026. URL https://arxiv.org/abs/2605.12408.

Hamrita Haridharan, Gomathy Dhanasekar, and Sharmila Nageswaran. Cognitive load gating system in motor imagery BCIs: A dual-task EEG study with differential entropy-based reliability estimation. *Frontiers in Artificial Intelligence*, 9:1859963, 2026. doi: 10.3389/frai.2026.1859963. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2026.1859963/full.

Vinay Jayaram and Alexandre Barachant. Moabb: Trustworthy algorithm benchmarking for brain-computer interfaces. *Journal of Neural Engineering*, 15(6):066011, 2018. doi: 10.1088/1741-2552/aadea0.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

Fabien Lotte, Harold Mouchère, and Anatole Lécuyer. Pattern rejection strategies for the design of self-paced EEG-based brain–computer interfaces. In *2008 19th International Conference on Pattern Recognition*, pp. 1–5, 2008. doi: 10.1109/ICPR.2008.4761454. URL https://ieeexplore.ieee.org/document/4761454.

Daily Milanés-Hermosilla, Rafael Trujillo Codorniú, René López-Baracaldo, Roberto Sagaró-Zamora, Denis Delisle-Rodriguez, John Jairo Villarejo-Mayor, and José Ricardo Núñez-Álvarez. Monte carlo dropout for uncertainty estimation and motor imagery classification. *Sensors*, 21(21):7241, 2021. doi: 10.3390/s21217241.

Daily Milanés-Hermosilla, Rafael Trujillo-Codorniú, Saddid Lamar-Carbonell, Roberto Sagaró-Zamora, Jorge Jadid Tamayo-Pacheco, John Jairo Villarejo-Mayor, and Denis Delisle-Rodriguez. Robust motor imagery tasks classification approach using Bayesian neural network. *Sensors*, 23(2):703, 2023. doi: 10.3390/s23020703.

Gerwin Schalk. Eeg motor movement/imagery dataset (version 1.0.0), 2009. URL https://physionet.org/content/eegmmidb/1.0.0/. Dataset.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas D. J. Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. doi: 10.1002/hbm.23730.

Naoki Tomida, Toshihisa Tanaka, Shunsuke Ono, Masao Yamagishi, and Hiroshi Higashi. Active data selection for motor imagery EEG classification. *IEEE Transactions on Biomedical Engineering*, 62(2):458–467, 2015. doi: 10.1109/TBME.2014.2358536.

Jose Antonio Uriguen and Begonya Garcia-Zapirain. Eeg artifact removal—state-of-the-art and guidelines. *Journal of Neural Engineering*, 12(3):031001, 2015. doi: 10.1088/1741-2560/12/3/031001.

Haoming Zhang, Mingqi Zhao, Chen Wei, Dante Mantini, Zherui Li, and Quanying Liu. Eegdenoisenet: a benchmark dataset for deep learning solutions of eeg denoising. *Journal of Neural Engineering*, 18(5):056057, 2021. doi: 10.1088/1741-2552/ac2bf8. URL https://github.com/ncclabsustech/EEGdenoiseNet.

A Sensitivity after refitting source models on a separate participant group

Table 5 shows unresolved clean (-0.058%) and EMG (-0.101%) direct effects. Its intervals fix the source fits and sham draws, omitting variation over source partitions and sham constructions.

Table 5: Sensitivity after refitting source models on a separate participant group for intact channels. Full-minus-sham normalized pAURC differences and paired 10,000-draw participant-bootstrap 95% intervals are in %; lower favors full. All 86 targets share the same channel layout; interactions use unrounded participant effects.

Targets	**Condition**	**Effect (%)↓**	**95% interval (%)**
86 common	Clean	-0.058	$[-0.233, +0.130]$
86 common	Synthetic EMG	-0.101	$[-0.346, +0.146]$
86 common	EMG − clean interaction	-0.043	$[-0.247, +0.159]$
89 original	Clean	-0.052	$[-0.225, +0.130]$
89 original	Synthetic EMG	-0.107	$[-0.351, +0.138]$
89 original	EMG − clean interaction	-0.054	$[-0.258, +0.145]$

Table 6: Exploratory differences between full-selector and sham pAURC under 50% channel loss ($n = 86$): effect and paired participant-bootstrap 95% interval (%; 10,000 draws), lower favors full. Each interval conditions on its saved source and target sham draws; seed provenance is stated in the uncertainty section.

Channel loss	**Trial**	**Original effect (%)↓**	**Disjoint effect (%)↓**
Independent	Clean	-0.051 $[-0.120, +0.013]$	-0.105 $[-0.162, -0.047]$
Independent	EMG	-0.101 $[-0.220, +0.012]$	-0.106 $[-0.203, -0.013]$
Independent	EOG	-0.037 $[-0.111, +0.033]$	-0.113 $[-0.170, -0.055]$
Persistent	Clean	$+0.049$ $[-0.029, +0.135]$	$+0.082$ $[-0.053, +0.224]$
Persistent	EMG	$+0.001$ $[-0.110, +0.115]$	-0.048 $[-0.185, +0.087]$
Persistent	EOG	$+0.014$ $[-0.048, +0.086]$	$+0.102$ $[-0.027, +0.242]$

Table 6 shows nominally lower pAURC for the full selector than for the sham on three trials under independent channel loss after refitting the source decoders. No such interval excludes zero under persistent channel loss. This exploratory grid is unadjusted for multiple comparisons and does not show whether the amplitude score helps more under contamination than on clean trials.

B Additional artifact and channel-loss conditions

The EMG and EOG perturbations were applied with intact channels and with independent or persistent 50% channel loss. Table 7 shows that every interaction interval includes zero; the table does not resolve the primary comparison.

Table 7: Changes from clean trials in the difference between full-selector and sham pAURC for 86 participants with a shared channel layout. Each entry is an EMG or EOG effect in %, with 10,000-draw paired participant-bootstrap 95% intervals. Negative values favor the full selector under the named artifact.

Channel state	**Synthetic EMG − clean (%)**	**Synthetic EOG − clean (%)**
Intact	-0.021 $[-0.129, +0.096]$	$+0.019$ $[-0.030, +0.070]$
Independent 50% loss	-0.050 $[-0.169, +0.062]$	$+0.014$ $[-0.034, +0.061]$
Persistent 50% loss	-0.047 $[-0.138, +0.049]$	-0.035 $[-0.083, +0.014]$

C Coverage profile of the exploratory EMG pattern

Table 8 shows how the EMG-related change in the full-selector minus sham pAURC difference varies by coverage. Intervals exclude zero when the selector accepts 10–40% of trials, include zero from 50% onward, and equal zero when every trial is accepted. These exploratory intervals are unadjusted across coverage values and do not establish a change in the originally specified pAURC endpoint.

Table 8: Exploratory EMG-related change from clean in the difference between full-selector and sham accepted-trial risk at each coverage ($n = 86$). Effects and 95% bootstrap intervals are %; negative values favor the full selector. This profile does not replace the originally specified pAURC analysis.

Coverage (%)	10	20	30	40
Effect (%)↓	-2.47	-2.13	-1.21	-0.41
95% interval	$[-3.62, -1.29]$	$[-2.94, -1.35]$	$[-1.76, -0.70]$	$[-0.72, -0.11]$
Coverage (%)	50	60	70	80
Effect (%)↓	-0.09	$+0.05$	-0.09	-0.05
95% interval	$[-0.42, +0.25]$	$[-0.18, +0.29]$	$[-0.25, +0.08]$	$[-0.21, +0.11]$
Coverage (%)	90	100		
Effect (%)↓	$+0.026$	0.000		
95% interval	$[-0.067, +0.121]$	$[0.000, 0.000]$		

D Initial channel-loss cohort

Table 9: Initial PhysioNetMI channel-loss diagnostic: 20 target participants, 50% independent test-channel loss, three seeds averaged within participant. The comparison of full and matched selectors changes the number of fitted coefficients.

Selector	**pAURC (%)↓**
Confidence	47.30
Matched	48.76
Full	49.06

E Earlier Cho2017 transfer results

This secondary analysis evaluates selective accuracy on Cho2017 with 50% test-channel dropout. For each target participant, a balanced logistic-regression model estimates whether decoder predictions are correct from 13 standardized trial descriptors and correctness labels from the other 51 participants. We compare confidence-only ranking with three alternatives: ranking by the model's predicted probability of correct decoding, averaging that rank equally with confidence rank, and averaging it equally with the stored quality-score rank. The available records do not establish whether this stored quality-score field is the calibration-referenced amplitude score used in the primary PhysioNet analysis.

Across the complete grid, the predicted-correctness rank and its quality-rank blend have positive participant-level 95% interval lower bounds in 12 of 12 settings; the confidence-rank blend does so in 1 of 12. The same 52 participants recur across 36 settings, so the intervals are not adjusted for these repeated comparisons. The quality blend is not directly compared with the predicted-correctness rank, so its added value over that ranking is unknown.

All 36 settings have files for all 52 target participants. In a zero-fill check, missing target files would contribute a zero difference. No target file is missing here, so this check leaves every displayed estimate and interval unchanged; it does not test sensitivity to new target participants or to a different stored quality score.

Table 10: Cho2017 transfer results for the mask-aware decoder under 50% test-channel dropout. Each table entry reports selective accuracy for the candidate policy minus selective accuracy for confidence-only ranking, as an absolute difference (%) with a paired Student-t 95% interval over 52 target participants. Positive values favor the candidate. The same participants recur across settings, so intervals are descriptive and are not adjusted across the grid.

Calibration trials	Coverage (%)	Candidate policy	Absolute difference (%)↑	95% interval (%)
8	80	Predicted correctness rank	$+2.17$	$[+1.74, +2.60]$
8	80	Correctness and confidence ranks	$+0.07$	$[-0.22, +0.35]$
8	80	Correctness and quality-score ranks	$+1.84$	$[+1.35, +2.34]$
8	90	Predicted correctness rank	$+1.16$	$[+0.87, +1.45]$
8	90	Correctness and confidence ranks	-0.22	$[-0.45, +0.01]$
8	90	Correctness and quality-score ranks	$+0.87$	$[+0.54, +1.19]$
32	80	Predicted correctness rank	$+2.02$	$[+1.47, +2.58]$
32	80	Correctness and confidence ranks	$+0.21$	$[-0.12, +0.53]$
32	80	Correctness and quality-score ranks	$+2.30$	$[+1.73, +2.87]$
32	90	Predicted correctness rank	$+1.34$	$[+0.98, +1.69]$
32	90	Correctness and confidence ranks	$+0.08$	$[-0.17, +0.33]$
32	90	Correctness and quality-score ranks	$+0.99$	$[+0.66, +1.32]$
64	80	Predicted correctness rank	$+2.64$	$[+2.19, +3.09]$
64	80	Correctness and confidence ranks	$+0.52$	$[+0.18, +0.85]$
64	80	Correctness and quality-score ranks	$+2.53$	$[+1.99, +3.06]$
64	90	Predicted correctness rank	$+1.19$	$[+0.86, +1.52]$
64	90	Correctness and confidence ranks	$+0.10$	$[-0.13, +0.33]$
64	90	Correctness and quality-score ranks	$+1.01$	$[+0.65, +1.37]$

Table 11: Cho2017 transfer results for the standard decoder under 50% test-channel dropout. Each table entry reports selective accuracy for the candidate policy minus selective accuracy for confidence-only ranking, as an absolute difference (%) with a paired Student-t 95% interval over 52 target participants. Positive values favor the candidate. The same participants recur across settings, so intervals are descriptive and are not adjusted across the grid.

Calibration trials	Coverage (%)	Candidate policy	Absolute difference (%)↑	95% interval (%)
8	80	Predicted correctness rank	$+2.58$	$[+2.07, +3.10]$
8	80	Correctness and confidence ranks	$+0.08$	$[-0.28, +0.45]$
8	80	Correctness and quality-score ranks	$+2.31$	$[+1.72, +2.90]$
8	90	Predicted correctness rank	$+1.16$	$[+0.80, +1.52]$
8	90	Correctness and confidence ranks	$+0.16$	$[-0.09, +0.41]$
8	90	Correctness and quality-score ranks	$+1.12$	$[+0.74, +1.50]$
32	80	Predicted correctness rank	$+2.44$	$[+1.92, +2.97]$
32	80	Correctness and confidence ranks	$+0.01$	$[-0.31, +0.32]$
32	80	Correctness and quality-score ranks	$+2.92$	$[+2.42, +3.42]$
32	90	Predicted correctness rank	$+1.41$	$[+1.08, +1.73]$
32	90	Correctness and confidence ranks	$+0.21$	$[-0.06, +0.48]$
32	90	Correctness and quality-score ranks	$+1.32$	$[+0.97, +1.67]$
64	80	Predicted correctness rank	$+2.60$	$[+2.01, +3.19]$
64	80	Correctness and confidence ranks	$+0.27$	$[-0.02, +0.56]$
64	80	Correctness and quality-score ranks	$+3.09$	$[+2.45, +3.73]$
64	90	Predicted correctness rank	$+1.40$	$[+1.03, +1.76]$
64	90	Correctness and confidence ranks	$+0.17$	$[-0.01, +0.36]$
64	90	Correctness and quality-score ranks	$+1.38$	$[+1.00, +1.75]$

An earlier 20-target snapshot evaluated quality-only ranking under the same channel-drop and artifact conditions at the eight-trial budget. None of its four 80% coverage intervals has a positive lower bound, and two lie wholly below zero. This smaller analysis uses a different selector and is shown as a boundary, not as a replication of the learned-rank grid.

Table 12: Earlier Cho2017 quality-only ranking under eight-trial calibration, 50% test-channel dropout, and synthetic EMG on 20 target participants. Entries are quality-only minus confidence-only selective accuracy at 80% coverage, in percent, with participant-level 95% intervals. Training channel-drop probabilities are 0.05 and 0.25. No interval has a positive lower bound; two lie wholly below zero. The same participants recur across settings, and the four intervals are not adjusted for multiplicity.

Decoder	**Train drop rate**	**Difference (%)↑**	**95% interval (%)**
Mask-aware	0.05	-0.4	$[-1.0, +0.3]$
Mask-aware	0.25	-0.9	$[-1.5, -0.3]$
Standard	0.05	$+0.0$	$[-0.5, +0.6]$
Standard	0.25	-0.7	$[-1.3, -0.1]$

F Exploratory secondary selective-risk metrics

Table 13 gives secondary intact-channel changes under synthetic EMG relative to clean in the full-versus-sham effects for $n = 86$. Confidence intervals for the secondary AURC and AUGRC outcomes lie below zero, but the grid spans multiple metrics and conditions without adjustment for multiple comparisons. These results do not overturn the unresolved pAURC comparison.

Table 13: Exploratory changes from clean under synthetic EMG in the difference between full and sham on intact channels ($n = 86$). Effects and participant-bootstrap 95% intervals are %; negative values favor the full selector.

Metric	**Effect (%)↓**	**95% interval (%)**
pAURC (50–100%)	-0.021	$[-0.129, +0.096]$
AURC	-0.673	$[-0.941, -0.423]$
AUGRC	-0.135	$[-0.201, -0.069]$
Risk at 80% coverage	-0.052	$[-0.207, +0.114]$
Risk at 90% coverage	$+0.026$	$[-0.067, +0.121]$
Risk at 100% coverage	0.000	$[0.000, 0.000]$