Supplementary Material S1

Conditional sensitivity of the time-aware versus permuted comparison

Manuscript: electronics-4596883. This supplement reports a retrospective reference-model sensitivity analysis. It does not report new benchmark fits or an estimate of the exact power of the complete benchmark.

S1.1. Reference model and test

For each coupling strength (0, 2, 4, 6), 50,000 samples of 20 paired ROC AUC differences were simulated. Each sample used independent, identically distributed symmetric differences with a common location shift of 0, 0.02, 0.04, 0.06, or 0.08. Normal innovations and Student t innovations with five degrees of freedom were used; the latter were multiplied by sqrt(3/5) to give unit variance. Both were then scaled by the observed sample standard deviation of the 20 time-aware-minus-permuted differences.

Coupling	Observed sample SD
0	0.06690924
2	0.04789570
4	0.05628079
6	0.07557748

The standard deviations pool four repetitions from each of five fixed acoustic configurations and use the sample definition (ddof = 1). They are treated as fixed reference values; their estimation uncertainty is not propagated.

S1.2. Meaning of the bounds

Each sample was tested with the exact two-sided Wilcoxon signed-rank test. In a 12-comparison Holm family at alpha = 0.05, $p \leq 0.05/12$ guarantees rejection and rejection requires $p \leq 0.05$. The simulated frequencies at these thresholds therefore estimate lower and upper bounds on the focal Holm rejection probability under the reference model. Effects in the other 11 comparisons need not be specified. These are **not confidence intervals**; Monte Carlo uncertainty is reported separately.

S1.3. Scope and interpretation

The reference model does not retain configuration-specific means or variances, introduce dependence between repetitions, or include uncertainty from new recordings, refitted acoustic classifiers, or farms. Sharing fixed empirical pools does not by itself establish stochastic dependence between separately generated realizations. Normal and t reference distributions are unbounded approximations to bounded AUC differences. Hypothetical shifts are not clinically established equivalence margins.

At a shift of 0.02, even the upper rejection-probability estimates were below 0.50 in every scenario. This indicates limited sensitivity to that shift under these assumptions. It does not establish why the actual primary comparisons were non-significant, or demonstrate equivalence.

Reproducibility: seed 20261031; NumPy 2.3.5; SciPy 1.17.0. Run python power_sensitivity.py with reference_scales.json in the same directory. Full numerical output is supplied as JSON and CSV.

S1.4. Normal reference distribution

All entries are probabilities (0 to 1), not percentages. Lower: rejection at 0.05/12. Upper: rejection at 0.05. Both are simulated estimates of bounds on Holm rejection probability under the stated reference model.

Coupling	Shift (AUC)	Lower	Upper	MCSE (lower)	MCSE (upper)
0	0.00	0.0038	0.0475	0.00028	0.00095
0	0.02	0.0420	0.2328	0.00090	0.00189
0	0.04	0.2872	0.6942	0.00202	0.00206
0	0.06	0.7259	0.9616	0.00199	0.00086
0	0.08	0.9607	0.9985	0.00087	0.00017
2	0.00	0.0038	0.0475	0.00028	0.00095
2	0.02	0.1027	0.4040	0.00136	0.00219
2	0.04	0.6399	0.9346	0.00215	0.00111
2	0.06	0.9763	0.9994	0.00068	0.00011
2	0.08	0.9997	1.0000	0.00007	0.00000
4	0.00	0.0038	0.0475	0.00028	0.00095
4	0.02	0.0667	0.3090	0.00112	0.00207
4	0.04	0.4491	0.8368	0.00222	0.00165
4	0.06	0.8955	0.9930	0.00137	0.00037
4	0.08	0.9958	1.0000	0.00029	0.00003
6	0.00	0.0038	0.0475	0.00028	0.00095
6	0.02	0.0314	0.1927	0.00078	0.00176
6	0.04	0.2044	0.5861	0.00180	0.00220
6	0.06	0.5782	0.9090	0.00221	0.00129
6	0.08	0.8905	0.9924	0.00140	0.00039

MCSE is the Monte Carlo standard error of each reported frequency: sqrt(p(1-p)/50,000). Its maximum possible value is approximately 0.00224. It measures simulation precision, not uncertainty in the empirical standard deviations or validity of the reference-model assumptions. Zero-shift rows provide a reference-null check.

Common random draws were reused across scales and shifts within each reference distribution. This does not model the joint outcomes of the benchmark comparisons, and Monte Carlo errors for different scenarios are not independent.

S1.5. Variance-standardized Student t(5) reference distribution

All entries are probabilities (0 to 1), not percentages. Lower: rejection at 0.05/12. Upper: rejection at 0.05. Both are simulated estimates of bounds on Holm rejection probability under the stated reference model.

Coupling	Shift (AUC)	Lower	Upper	MCSE (lower)	MCSE (upper)
0	0.00	0.0039	0.0503	0.00028	0.00098
0	0.02	0.0612	0.2894	0.00107	0.00203
0	0.04	0.3855	0.7774	0.00218	0.00186
0	0.06	0.7888	0.9732	0.00183	0.00072
0	0.08	0.9550	0.9983	0.00093	0.00018
2	0.00	0.0039	0.0503	0.00028	0.00098
2	0.02	0.1509	0.4934	0.00160	0.00224
2	0.04	0.7230	0.9559	0.00200	0.00092
2	0.06	0.9690	0.9990	0.00078	0.00014
2	0.08	0.9973	1.0000	0.00023	0.00002
4	0.00	0.0039	0.0503	0.00028	0.00098
4	0.02	0.0970	0.3825	0.00132	0.00217
4	0.04	0.5577	0.8880	0.00222	0.00141
4	0.06	0.9091	0.9937	0.00129	0.00035
4	0.08	0.9892	0.9998	0.00046	0.00006
6	0.00	0.0039	0.0503	0.00028	0.00098
6	0.02	0.0446	0.2366	0.00092	0.00190
6	0.04	0.2841	0.6821	0.00202	0.00208
6	0.06	0.6727	0.9384	0.00210	0.00108
6	0.08	0.9057	0.9935	0.00131	0.00036

MCSE is the Monte Carlo standard error of each reported frequency: sqrt(p(1-p)/50,000). Its maximum possible value is approximately 0.00224. It measures simulation precision, not uncertainty in the empirical standard deviations or validity of the reference-model assumptions. Zero-shift rows provide a reference-null check.

Common random draws were reused across scales and shifts within each reference distribution. This does not model the joint outcomes of the benchmark comparisons, and Monte Carlo errors for different scenarios are not independent.

Statistical references: Wilcoxon, F. Individual Comparisons by Ranking Methods. Biometrics Bulletin 1945, 1, 80-83. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics 1979, 6, 65-70.