TRAINING SCORES FOR CONDITIONAL AUDITS: OVERLAP WEIGHTING AND POWER UNDER CONFOUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional audits often test a score trained for marginal prediction, even though that objective rewards the very confounders the audit conditions on. We study whether train-only overlap weighting can recover conditional signal lost during score construction. The procedure estimates discrete-stratum label proportions in training data, fits a weighted classifier, and freezes its scores before an independent exact-stratum permutation test. We establish its training-balance target and explain why independent testing preserves the randomization guarantee. A Gaussian power ceiling separates score-learning loss from limited overlap. Across 25 synthetic configurations, we evaluate effect size, overlap, sample size, noise, and confounder measurement error. In a fresh 500-seed alternative block, overlap weighting raises forest detection from 61.6% to 79.4%, a paired gain of 17.8 percentage points [13.4, 22.4], while its null rejection rate is 4.0%. Global class balancing yields 60.8% detection. Additional continuous-confound and repeated-clip experiments delineate the test's assumptions. The contribution is a controlled score-training study and reproducible audit procedure built from established weighting and randomization tools.

1 INTRODUCTION

A model trained to predict a label can become a poor instrument for testing whether a representation contains information beyond known confounders. Marginal prediction rewards demographic and acquisition shortcuts. A conditional audit later asks the fitted score to distinguish labels within groups where those shortcuts are fixed. If training has compressed a weak conditional signal out of the score, a valid test can have little power even when the representation contains usable information.

This distinction is actionable. Changing the score-training objective need not change the held-out test or its assumptions. We study a simple intervention: estimate label proportions within prespecified training strata, fit a classifier with bounded overlap weights, and freeze its scores before conditional testing on an independent sample. The weights equalize the two classes within training strata and concentrate the loss on regions where both labels occur. They use no test labels and require no inverse-propensity division.

Overlap weights and conditional randomization are established tools. The question addressed here is their interaction at the score-construction stage: can training for the overlap population recover conditional-test power lost by a classifier optimized for marginal discrimination? We isolate this question with matched datasets, learner settings, and randomization draws. Ordinary class balancing, a second classifier family, and a known signal-channel oracle distinguish within-stratum balancing from generic weighting and from information absent in the features.

Our evidence combines a 25-configuration synthetic grid with a fresh-seed validation block. In that block, overlap weighting increases forest power from 61.6% to 79.4% at an injected effect of 0.5, with a paired gain of 17.8 percentage points (95% seed-bootstrap interval 13.4–22.4). The weighted forest's null rejection rate is 4.0%. Its lower marginal AUROC illustrates that a score suitable for prediction and one suitable for a conditional audit need not be the same. The experiments also quantify how mismeasured covariates, continuous binning, and repeated clips affect the testing step.

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.

The paper makes three contributions. First, it provides a reproducible, train-only overlap-weighted scoring procedure evaluated specifically for conditional-audit power. Second, it separates score-learning loss from design information through a Gaussian power ceiling and a signal-channel oracle. Third, it supplies a complete stress-test suite and a discrete control-allocation calculation, linking observed overlap to an implementable training or sampling intervention. The motivating facial-video example is used for aggregate arithmetic and a paired-world reproduction; the claims of the new method are statistical and synthetic, not clinical.

2 Related work and scope

Shortcut learning can undermine transfer beyond standard benchmarks (Geirhos et al., 2020); in medical imaging, demographic and acquisition variables are important sources. Jabbour et al. (2020) demonstrate demographic prediction and shortcut exploitation in chest radiographs. Makar et al. (2022) study causally motivated shortcut removal using auxiliary labels. Our question precedes removal: whether an evaluation can distinguish a confound-only mechanism from an additional label-dependent component when relevant support is sparse.

Conditional independence testing requires assumptions. Model-X methods exploit conditional distribution information (Candès et al., 2018), while classifier-based testing turns dependence detection into prediction (Sen et al., 2017). Berrett et al. (2020) develop a conditional permutation test based on a conditional distribution model; it is not interchangeable with arbitrary permutation within coarse bins. Shah & Peters (2020) establish fundamental limitations of general conditional independence testing. We use only a simpler finite-stratum randomization argument, explicitly requiring within-stratum exchangeability and independent training. We make no distribution-free claim for continuous, coarsened, or unmeasured confounders. Holdout randomization already uses independent fitting to make conditional tests practical (Tansey et al., 2021); our use of sample splitting is not a new test. Distilled randomization (Liu et al., 2020) and weighted residual-covariance statistics (Scheidegger et al., 2022) offer other routes to computational efficiency or power. Our overlap weights act on the training loss, not on a residual-covariance test statistic. Spisak (2022) studies related confound tests on predictions. Its partial-confound question, whether a score remains associated with a confounder conditional on the outcome, differs from our question of score–label association conditional on measured confounders.

Propensity scores formalize covariate balancing (Rosenbaum & Rubin, 1983), and limited overlap motivates restricting the inferential target (Crump et al., 2009). Overlap weighting targets a balanced population (Li et al., 2018). We use these established weights to train an audit score rather than claim a new weighting estimator. Test-statistic design can materially affect conditional-testing power (Sudarshan et al., 2021); our comparison concerns the same exact-stratum test with different train-only scores, not superiority over all conditional-independence tests.

The reference feature representation uses isometric log-ratio (ILR) coordinates, an established construction for compositions (Egozcue et al., 2003). It is included to make a future motion experiment implementable, not as a new representation-learning algorithm. Our synthetic summary features are not generated by running a landmark detector on video. Thus neither synthetic AUROC nor a synthetic label effect is evidence about a biological dementia mechanism.

3 Motivating example and statistical target

An aggregate facial-video screening artifact reports 150 positive-provenance and 150 community-control clips. Its age-bin counts imply an age-bin AUROC of 0.921 and at most 41 disjoint bin-matched pairs. Our paired-world reproduction gives median feature AUROC 0.960 without an added label effect and 0.969 with one. These exact-count and synthetic observations motivate the conditional question; the full arithmetic and reproduction are in Supplement S12. They are not patient-level validation.

The operational target is whether an independently trained score $s(X)$ remains associated with the observed label Y conditional on declared covariates Z. This differs from marginal prediction, from independence of the score and Z conditional on Y, and from identifying a causal disease effect. A

score can fail to retain information present in X, so a valid non-rejection for $s(X)$ need not imply conditional independence of the full representation.

4 SEPARATE SCORE LOSS FROM DESIGN INFORMATION

Let Y denote the observed binary label, Z measured confounders, U unmeasured acquisition variables, and X the measured representation. A confound-only mechanism has the form

$$X = h(Z, U, \varepsilon), \tag{1}$$

where Y is associated with (Z, U) through sampling or label provenance. An alternative mechanism adds a label-dependent component. This is a statistical construction: replacing Y by a clinical disease variable would require further measurement and causal assumptions.

Proposition 1 (No separation on deterministic-label strata)**.** *Suppose $Y = g(Z)$ almost surely on a subset of the observed support. On that subset, the models $X = h(Z) + \delta Y + \varepsilon$ and $X = \widetilde{h}(Z) + \varepsilon$, where $\widetilde{h}(z) = h(z) + \delta g(z)$, induce the same observed distribution. No statistic of the observed data distinguishes these mechanisms there.*

Proof. Substitution of $Y = g(Z)$ makes the two equations identical for every observation on that support. Identical data distributions induce identical distributions for every measurable statistic. □

The proposition concerns exact non-overlap. It does not say that an age gap alone makes every effect unidentifiable, or that no useful inference is possible in overlapping strata. Partial overlap instead restricts the population over which a conditional contrast is supported and can sharply reduce precision. Extrapolation outside that population requires assumptions that should be stated rather than hidden inside a fitted model.

Proposition 2 (A finite-stratum overlap diagnostic)**.** *For fixed stratum counts n_{1z}, n_{0z} and $n_z = n_{1z} + n_{0z}$, the sum of squared residualized labels is*

$$I_Y = \sum_i (Y_i - \overline{Y}_{z_i})^2 = \sum_z \frac{n_{1z} n_{0z}}{n_z}. \tag{2}$$

Single-class strata contribute zero. Under the fixed-design model $X_i = \alpha_{z_i} + \delta Y_i + \varepsilon_i$ with independent homoskedastic errors of variance σ^2, the least-squares estimate of δ has variance σ^2 / I_Y, when $I_Y > 0$.

Proof. Within a stratum, there are n_{1z} residuals $1 - n_{1z}/n_z$ and n_{0z} residuals $-n_{1z}/n_z$. Expanding their squared sum gives $n_{1z} n_{0z}/n_z$. Regressing out the stratum intercepts leaves the scalar regressor $Y_i - \overline{Y}_{z_i}$; the usual fixed-design least-squares variance follows. □

The cross-class pair count $\sum_z n_{1z} n_{0z}$ is not an independent sample size: the same participant occurs in many pairs. Equation 2 is an information diagnostic under the stated linear error model, not a universal effective sample size for AUROC. It nevertheless exposes why additional single-class observations cannot repair missing within-stratum comparisons.

Audits answer different questions. A metadata baseline tests whether a particular fitted metadata predictor works. Matching tests a restricted and reweighted population. Residualizing X by an estimate of $\mathbb{E}[X \mid Z]$ removes a conditional mean, not necessarily conditional distributional dependence. An age-prediction diagnostic within controls detects an age pathway but does not prove that all label prediction uses that pathway. Each is informative within its scope; none alone identifies a disease-specific mechanism.

4.1 A QUANTITATIVE LIMIT ON POWER AND A RECRUITMENT CONSEQUENCE

The deterministic-stratum result describes complete failure of identification. To quantify partial overlap, consider a scalar Gaussian measurement conditional on fixed observed labels and strata,

$$X_i = \alpha_{z_i} + \delta Y_i + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2). \tag{3}$$

The stratum means α_z are unrestricted nuisance parameters; σ is known. This restrictive model gives an interpretable best-case bound, not a characterization of arbitrary video embeddings.

Proposition 3 (Gaussian conditional power ceiling)**.** *For a fixed design (Y, Z) and positive effect δ, any test with size at most α for every null stratum-mean vector has power at most*

$$\Phi\left(\frac{\delta\sqrt{I_Y}}{\sigma} - \Phi^{-1}(1-\alpha)\right). \tag{4}$$

For known σ, this bound is attained by the one-sided Gaussian test based on $\sum_i (Y_i - \overline{Y}_{z_i})X_i$. In particular, $I_Y = 0$ reduces the bound to α.

Proof. Compare the alternative mean $\alpha_{z_i} + \delta Y_i$ with the null mean $\alpha_{z_i} + \delta\overline{Y}_{z_i}$. Their difference is $\delta(Y_i - \overline{Y}_{z_i})$ and its squared Euclidean norm is $\delta^2 I_Y$. The likelihood-ratio test between these two Gaussian distributions has the power in Equation 4. Any uniformly level-α test must also control size at this particular null, so the Neyman–Pearson bound applies. The residual-label contrast cancels every stratum intercept and has variance $\sigma^2 I_Y$, achieving that bound. □

This is an application of Gaussian testing theory. Its value here is to distinguish a learner's failure to extract signal from a design that contains too little information even under favorable assumptions. In the aggregate age-bin design, $I_Y = 29.535$, compared with 75 for 300 observations balanced within strata. At $\delta/\sigma = 0.3$ and $\alpha = 0.05$, the corresponding ceilings are 0.494 and 0.830. These are analytic Gaussian quantities, not estimated power for dementia screening. Adding recording-style strata can only reduce I_Y for a fixed sample, because conditioning on a finer partition reduces within-stratum label variation.

Where should another control be collected? If a stratum has a positive and b control observations, adding one independent control increases its information by

$$\Delta I = \frac{a^2}{(a+b)(a+b+1)}. \tag{5}$$

For fixed positive counts and equal recruitment costs, these gains decrease as controls are added. Greedily selecting the largest remaining gain therefore maximizes total I_Y for an integer control budget; an exchange argument is given in Supplement S2. This rule prioritizes missing support instead of simply increasing the total sample size. Applied to the printed age bins alone, 100 optimally allocated additional controls raise I_Y to 73.619. This calculation assumes accessible, independent controls in the stated strata. Unknown recording-style counts, differential costs, and target-population constraints can change the allocation; it is a planning illustration rather than a recommendation that 100 controls suffice clinically.

5 Train the score for the conditional question

The actionable gap is a mismatch between training and testing objectives. Standard supervised training rewards every feature that predicts Y, including acquisition and demographic shortcuts. A conditional audit subsequently asks whether a one-dimensional score retains label information *after* those shortcuts are accounted for. Compressing X to a score optimized for marginal prediction can therefore weaken the audit even when a useful conditional signal exists. The oracle comparison in our controlled generator measures that avoidable loss.

Train-only overlap weighting. For a prespecified discrete confounder stratum z, estimate $\widehat{e}_z = n_{1z}^{\mathrm{tr}}/n_z^{\mathrm{tr}}$ from training data only. Give training observation i weight

$$w_i = Y_i(1 - \widehat{e}_{z_i}) + (1 - Y_i)\widehat{e}_{z_i}, \qquad \widehat{s} \in \arg\min_s \sum_{i \in \mathrm{tr}} w_i \ell(Y_i, s(X_i)). \tag{6}$$

We normalize nonzero total weights to mean one before fitting, retaining the same regularization settings across comparators. These are established overlap weights (Li et al., 2018); our contribution is their controlled evaluation for learning a score used in an independent conditional audit. Unlike inverse-propensity weights, the unnormalized weights are bounded by one. A single-class training stratum has zero weight. If all strata are single-class, the procedure returns "no training overlap" instead of a fitted score.

Proposition 4 (Exact training balance and its population target)**.** *In every mixed-label training stratum, the total unnormalized weight of each class equals $n_{1z}^{\mathrm{tr}} n_{0z}^{\mathrm{tr}}/n_z^{\mathrm{tr}}$. With the true propensity $e(z)$, the population weighted distribution has $P_w(Y = 1 \mid Z = z) = 1/2$ wherever $0 < e(z) < 1$, and its stratum density is proportional to $P(Z = z)e(z)(1 - e(z))$.*

Proof. Positive weight is $n_{1z}(1 - n_{1z}/n_z) = n_{1z}n_{0z}/n_z$; negative weight is $n_{0z}n_{1z}/n_z$. In the population, both class-specific masses at z equal $P(z)e(z)(1 - e(z))$. Normalization gives the stated result. □

Balance removes the training reward for a predictor that uses only the declared strata. It does not force a finite-sample learner to ignore every proxy, and it does not balance omitted variables. The weighted training target emphasizes the overlap population. Test observations retain their original distribution: weights are used to learn the score, not to alter held-out labels, strata, or the randomization distribution.

Why global class balancing is not enough. Multiplying every positive by the same factor and every negative by another makes global class totals equal, but generally leaves different label propensities across strata. We therefore include class-balanced forests as an ablation, using the same trees, seeds, and held-out randomization banks. An improvement over this comparator isolates the effect of within-stratum balancing from ordinary class reweighting.

Validity is inherited from the independent test. The weights, feature scaling, classifier, and score orientation are all fixed before reading held-out labels. Conditional on training data, the resulting score is a fixed function of test features. Under the exact-stratum null, its test labels are exchangeable within strata, so changing the training loss changes power but not the finite-sample randomization argument. This is not a new universal conditional-independence test. It is a way to improve the score supplied to a valid test under declared assumptions.

Procedure: overlap-trained conditional audit. (1) Split independent units into training and test sets; fix strata and the positive score direction. (2) Compute training stratum counts and Equation 6; stop if no weights are positive. (3) Fit the weighted score using training data only. (4) Freeze scores on test observations; report test support and I_Y. (5) Permute labels within exact test strata, keeping scores fixed, and report the Monte Carlo p-value with the number of randomizations. (6) Interpret non-rejection jointly with the power benchmark and support diagnostics.

6 A CONDITIONAL TEST WITH EXPLICIT EXCHANGEABILITY

Separate fitting from testing. Let a scoring rule s be fitted on an independent training sample. In a held-out test sample, fix exact discrete strata $Z = z$. Under $H_0 : X \perp Y \mid Z$, independent observations, and identical label probabilities within each stratum, test labels are exchangeable within stratum conditional on their counts and the observed features. For example, use the rank statistic

$$T = \sum_i Y_i r_i, \qquad r_i = \mathrm{rank}(s(X_i)), \tag{7}$$

with midranks for ties. Draw B independent uniform within-stratum permutations and calculate

$$p = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{T_b \geq T\}}{B + 1}. \tag{8}$$

Conditional exchangeability gives a super-uniform randomization p-value; the plus-one correction avoids zero Monte Carlo values (Phipson & Smyth, 2010). Scores must not be selected using test labels. Cross-fitting alone does not make the entire vector of out-of-fold scores independent of all labels: another observation's label can influence a score. Holding that score vector fixed while permuting every label therefore needs a separate justification or a refitted randomization scheme.

What the guarantee does not cover. Binning continuous age or style does not generally yield identical label probabilities or conditional independence inside each bin. Adding a style variable can improve an empirical rejection rate without proving validity. A test conditioning on age and glasses also tests a different null from one conditioning on age, glasses, and style. Under a purely style-mediated mechanism, rejection of the former can be statistically correct while still being a false attribution to disease. We call this *omitted-confound rejection*, rather than automatically calling it a Type I error for the smaller conditioning set.

7 Robustness beyond the motivating generator

A single synthetic counterexample disproves a universal claim but gives little guidance about prevalence or severity of failures. We therefore extend the experiment along five axes and retain every declared configuration. The suite includes 25 discrete configurations, each with 200 independent seed replicates; two continuous-confound conditions with 500 replicates each; and six repeated-observation conditions with 500 replicates each. These are descriptive stress tests, not preregistered clinical analyses. The configuration grid was fixed before these extended runs, after the initial exploratory experiments.

Common protocol and comparators. Draw independent $A \sim \mathrm{Uniform}\{0,1,2\}$, $G \sim \mathrm{Bernoulli}(0.3)$ and $S \sim \mathrm{Bernoulli}(0.5)$. Labels have propensity $\mathrm{logit}^{-1}\{\lambda[1.8(A-1)+0.8G+2.4(S-0.5)]\}$. The six features are $(A, 1.5S, G, (A-1)^2, AS, \delta Y)$ plus independent $\mathcal{N}(0,\sigma^2)$ noise. Exact strata are (A, G, S). We vary $\delta \in \{0, 0.25, 0.5, 0.75, 1\}$, $\lambda \in \{0, 0.5, 1, 2, 4\}$, training and test sizes $n \in \{150, 300, 600\}$, feature-noise standard deviation $\sigma \in \{0.4, 0.8, 1.6\}$, and style-label flip probability $\eta \in \{0, 0.1, 0.25\}$. Training and test samples each contain n independent observations. This is a one-factor-at-a-time grid around $n = 300$, $\lambda = 1$, $\sigma = 0.8$, $\eta = 0$, including both null and alternative where relevant; it is not the full Cartesian product.

Each dataset is evaluated with an 80-tree random forest (Breiman, 2001) (minimum leaf size five), standardized logistic regression ($C = 1$), the known sixth feature as an oracle signal score, and the true propensity as a metadata-only negative control. All learned scores use independent training observations. Scores share permutation banks within each replicate, with 399 permutations and one-sided $\alpha = 0.05$. The oracle knows the location of the injected effect and is not a deployable baseline. The metadata-only score is constant within exact strata and must return $p = 1$, a useful implementation check.

Table 1: Central extended setting: $n = 300$, $\lambda = 1$, noise 0.8, $\delta = 0$ or 0.5. Each rejection rate uses 200 trials; brackets are pointwise exact binomial 95% intervals. AUROC is a replicate median.

Score	Null AUROC	Null rejection (%)	Power (%)
Random forest	0.753	4.0 [1.7, 7.7]	61.0 [53.9, 67.8]
Logistic regression	0.766	6.0 [3.1, 10.2]	79.0 [72.7, 84.4]
Oracle signal	–	5.5 [2.8, 9.6]	99.0 [96.4, 99.9]
Metadata only	–	0.0 [0.0, 1.8]	0.0 [0.0, 1.8]

Validity and power must be separated. At the central setting, the full-stratum forest test rejects in 8 null runs and 122 added-effect runs out of 200; logistic regression rejects in 12 and 158, respectively. The oracle rejects in 198 alternative runs. A learner can retain high marginal discrimination while losing conditional sensitivity because its score emphasizes confounders. Conversely, better conditional sensitivity need not imply higher marginal AUROC. All rates and exact binomial intervals are reported in the supplement, including configurations where a learner underperforms.

7.1 Fresh-seed validation of overlap-trained scores

After choosing the weighting rule from the exploratory results, we fixed the central generator, learner settings, strata, and positive score orientation and ran 500 new seeds in each world. These seeds are disjoint from the exploratory grid. The comparison therefore tests the selected training interven-

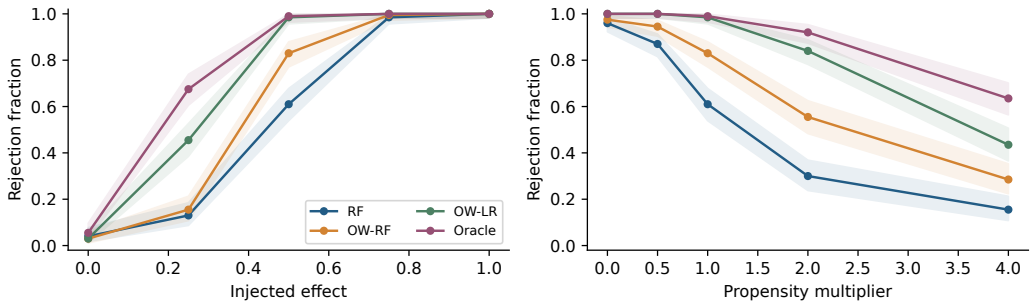

Figure 1: Extended discrete experiments. Conditional power is shown for standard, overlap-weighted, and oracle scores as the injected effect and the propensity multiplier change. Bands are pointwise exact binomial 95% intervals across 200 seeds; they are not simultaneous confidence bands. Null results and all settings appear in the separate supplement.

tion on new random realizations without selecting hyperparameters from this block. It remains a validation within the same synthetic family.

Table 2: Fresh seeds: 500 null and 500 alternative runs, effect 0.5. Rejection rates have pointwise exact 95% binomial intervals; AUROC is the median across seeds.

Score	Null rejection (%)	Power (%)	Alt. AUROC
Standard forest	6.8 [4.8, 9.4]	61.6 [57.2, 65.9]	0.790
Class-balanced forest	5.4 [3.6, 7.8]	60.8 [56.4, 65.1]	0.790
Overlap-weighted forest	4.0 [2.5, 6.1]	79.4 [75.6, 82.9]	0.660
Standard logistic	5.4 [3.6, 7.8]	72.4 [68.3, 76.3]	0.803
Overlap-weighted logistic	7.8 [5.6, 10.5]	95.2 [92.9, 96.9]	0.667
Oracle signal	5.2 [3.4, 7.5]	99.0 [97.7, 99.7]	–

The overlap-weighted forest detects the effect in 397/500 trials, compared with 308/500 for the standard forest and 304/500 for the globally class-balanced forest. The paired gain over the standard forest is 17.8 percentage points, with a 95% percentile interval of 13.4–22.4 obtained by resampling dataset seeds 10,000 times. Of the 500 paired alternative trials, 117 reject only with the overlap-weighted forest and 28 only with the standard forest. This improvement cannot be attributed to more test observations, a different classifier size, or a different permutation budget.

Under the alternative, median marginal AUROC decreases from 0.790 to 0.660 while conditional-test power increases. Weighting directs learning away from between-stratum prevalence differences; the held-out test rewards within-stratum ordering. These distinct metrics should not be combined into a claim that weighting improves every prediction task. The method is intended to construct an audit score, not to replace a calibrated deployment predictor without further evaluation.

Logistic regression offers a complementary check: power rises from 72.4% to 95.2%, compared with 99.0% for the oracle signal coordinate. The weighted logistic null rate is 7.8% in this 500-seed block, above the nominal level and explicitly retained in the table. A separate fixed-budget check on 5,000 further seeds gives 5.02% [4.43, 5.66]; both blocks are retained in Supplement S4, without retuning. Exact-stratum validity is a property of the independent randomization argument, while finite Monte Carlo fractions are noisy estimates that warrant checking when they depart from nominal behavior.

Overlap, sample size, and noise. Increasing λ makes the label increasingly predictable from confounders and changes the within-stratum information. The sample-size sweep increases both training and test observations; it therefore combines improvements in learning with improvements in test information. It must not be interpreted as a pure test-size experiment. The noise sweep changes the signal-to-noise ratio of every feature. We report these interventions separately rather

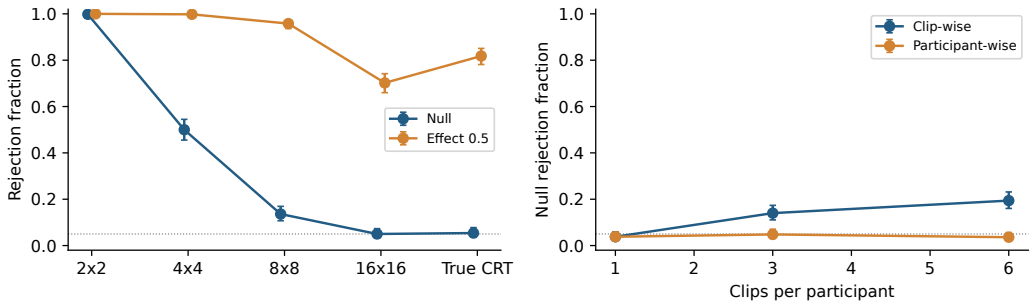

Figure 2: Two departures from the ideal test. Left: continuous covariates are coarsened before permutation, or handled through the true conditional law. Right: multiple clips share a participant label and random effect; participant-level and clip-level randomization answer different dependence assumptions. Error bars are pointwise exact binomial 95% intervals across 500 seeds.

than attributing all differences to overlap. The analytic ceiling in Equation 4 uses fixed design and a scalar Gaussian signal; the learned-score curves are not claimed to attain it.

Errors in recorded confounders. A measured recording-style label can be incomplete even when a column named "style" is present. We independently flip its binary annotation with probability η while preserving true style in the feature generator. Tests stratified by the corrupted label no longer condition on the variable that generated the confounding. At $\eta = 0.25$, the forest's null-world rejection fraction is 91.0%, compared with 4.0% when the true strata are used. These are failures of interpreting an incomplete conditioning set as complete control; they are not counterexamples to the exact-stratum theorem.

7.2 Continuous confounders: binning versus a known conditional law

Let A, S be independent standard Gaussians, $Y \mid A, S \sim \text{Bernoulli}(\text{logit}^{-1}(1.5A + 1.5S))$, and use the fixed score $V = A + S + \delta Y + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 1)$. This score has no fitted parameters or training-label reuse. We compare uniform permutations within $b \times b$ marginal-normal quantile bins, $b \in \{2, 4, 8, 16\}$, with conditional randomization drawing fresh labels from the known propensity. A second randomization incorrectly omits style from that propensity. All methods use 399 draws and 500 independent replicates in each world.

The known-propensity randomization is an oracle diagnostic of the test mechanism, not an estimated-propensity method. Under the null it rejects in 27 of 500 trials. Two-by-two bins reject in 499 trials. Finer bins trade residual confounding against increasingly sparse comparisons; full results include both rejection fractions and within-bin information. A small null rejection fraction in a fine partition can reflect conservatism and low information rather than a generally valid approximation.

A useful sensitivity statement. Let P be the true conditional test-label law given the fixed score inputs and covariates, and Q the law used for randomization. If the procedure is level α under Q, its rejection probability under P is at most $\alpha + \text{TV}(P, Q)$. This follows directly by applying total variation to the randomized rejection event. For independent Bernoulli labels with true and supplied probabilities p_i, q_i, a coupling gives $\text{TV}(P, Q) \le \min(1, \sum_i |p_i - q_i|)$. This conservative bound emphasizes that seemingly small conditional-model errors can accumulate; it does not estimate the error from data or justify arbitrary binning.

7.3 Repeated clips: the permutation unit matters

The final suite generates 100 independent participants with a four-level stratum, one label per participant, a shared Gaussian random effect, and $m \in \{1, 3, 6\}$ noisy clips per participant. A fixed clip score is stratum index plus δY plus the shared effect plus independent Gaussian noise of standard deviation 0.3. We compare clip-wise label permutation with permutation of participant labels

after averaging their clip scores. The latter respects the exchangeability unit; the former breaks within-participant dependence, a problem addressed by block permutation (Winkler et al., 2015).

For six clips per participant under the null, naive clip-wise testing rejects in 97 of 500 trials, versus 18 for participant-wise testing. This experiment holds the number of independent participants fixed: adding clips is not equivalent to recruiting more participants. The result concerns a simple constant-label, equal-cluster-size design. In a real corpus, sources can connect multiple identities, and connected-component grouping or a specifically justified cluster randomization may be necessary.

8 Implementation and practical use

The package exposes train-only weights, permutation banks, and support diagnostics, using scikit-learn estimators (Pedregosa et al., 2011). Rejection-rate intervals use the exact binomial construction of Clopper & Pearson (1934). All learned-score comparisons share test records and permutation draws within a seed. Weights use empirical training-stratum proportions; no true propensity is used to train the proposed score. Forests use 80 trees and minimum leaf size five. Logistic models use training-fitted standardization and $C = 1$. The oracle signal coordinate and true-propensity randomization appear only as diagnostic references.

Three outputs should accompany the test result: the score's marginal AUROC, the exact-stratum rejection result, and the amount of within-stratum information. Reporting only one confuses prediction, statistical validity, and sensitivity. We additionally report the conventional training-weight diagnostic $(\sum_i w_i)^2 / \sum_i w_i^2$; this is not substituted for test information or an independent sample count.

For repeated measurements, define the independent unit before splitting, training, and randomizing. For continuous covariates, the exact-stratum implementation should be replaced by a justified conditional-law method rather than silently binning. The accompanying continuous and cluster experiments demonstrate these boundaries. Hyperparameters and strata must not be selected to minimize the test p-value; an additional development split or a selection-aware test is needed for such choices.

The recruitment calculation complements training rather than replacing it. Weighting reallocates attention among observed training comparisons; it cannot create new independent comparisons in the test set. When test information is inadequate under a prespecified effect/noise model, Equation 5 identifies which eligible control strata contribute most information per equal-cost observation. This converts the overlap diagnosis into a concrete design calculation.

9 Limitations and conclusion

Overlap weighting is an established tool; our contribution is the score-training application, its connection to audit power, and the controlled evaluation. The numerical results establish behavior in declared synthetic families, not universal dominance or clinical validity. The weighting rule targets measured strata, and estimated continuous propensities require additional analysis. Conditional-score testing can miss dependence lost during score compression. Both the Gaussian power ceiling and the equal-cost recruitment rule rely on their stated models.

The experiments separate a remediable loss of conditional signal during training from limits imposed by overlap and test validity. Train-only overlap weighting offers a simple intervention at the first stage while retaining the independently held-out test. Its benefit must be judged by conditional sensitivity and calibrated null behavior, not by marginal AUROC alone.

ETHICS STATEMENT

This revision analyzes supplied aggregate counts and synthetic data. It does not release face images, public-video links, identity embeddings, participant-level health labels, or a deployable screening model. No new clinical study or video collection was conducted for these experiments. Documentation of consent or an institutional ethics determination for the original collection was not supplied; no approval, exemption, or consent is asserted here. Removing identifiable figures does not retrospectively establish permission for the original collection. Any human-data evaluation or release requires the responsible authors to establish the applicable institutional determination, rights, consent basis, and privacy safeguards. The synthetic findings should not be used to infer any individual's health status.

REPRODUCIBILITY STATEMENT

The anonymous supplement contains executable generators, per-replicate results, aggregate inputs, generated tables, and dependency versions. Seeds and all model settings are stated above. The main results can be regenerated without access to human records. Reference feature and evaluation utilities are provided with scope limitations; they are not represented as executed clinical experiments.

AI USE STATEMENT

Generative AI assisted with editing the manuscript and preparing the submission files following the venue format. The authors are responsible for the correctness, originality, provenance, and declarations of the final work.

REFERENCES

Thomas B. Berrett, Yi Wang, Rina Foygel Barber, and Richard J. Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B*, 82(1):175–197, 2020. doi: 10.1111/rssb.12340.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-X knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577, 2018. doi: 10.1111/rssb.12265.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96 (1):187–199, 2009. URL https://dash.harvard.edu/entities/publication/73120378-86bc-6bd4-e053-0100007fdf3b.

Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3): 279–300, 2003. doi: 10.1023/A:1023818214614.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Sarah Jabbour, David Fouhey, Ella Kazerooni, Michael W. Sjoding, and Jenna Wiens. Deep learning applied to chest x-rays: Exploiting and preventing shortcuts. In *Proceedings of Machine Learning for Healthcare*, volume 126, pp. 750–782, 2020. URL https://proceedings.mlr.press/v126/jabbour20a.html.

Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018. doi: 10.1080/01621459.2016.1260466.

Molei Liu, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas. Fast and powerful conditional randomization testing via distillation. Author manuscript, 2020. URL https://lucasjanson.fas.harvard.edu/papers/Fast_And_Powerful_Conditional_Randomization_Testing_Via_Distillation-Liu_Janson-2020.pdf.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 739–766, 2022. URL https://proceedings.mlr.press/v151/makar22a.html.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://www.jmlr.org/papers/v12/pedregosa11a.html.

Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010. doi: 10.2202/1544-6115.1585.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Cyrill Scheidegger, Julia Hörrmann, and Peter Bühlmann. The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273):1–68, 2022. URL https://jmlr.org/papers/v23/21-1328.html.

Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G. Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/02f039058bd48307e6f653a2005c9dd2-Abstract.html.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020. doi: 10.1214/19-AOS1857.

Tamas Spisak. Statistical quantification of confounding bias in machine learning models. *GigaScience*, 11:giac082, 2022. doi: 10.1093/gigascience/giac082.

Mukund Sudarshan, Aahlad Puli, Lakshmi Subramanian, Sriram Sankararaman, and Rajesh Ranganath. Contra: Contrarian statistics for controlled variable selection. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 1900–1908, 2021. URL https://proceedings.mlr.press/v130/sudarshan21a.html.

Wesley Tansey, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M. Blei. The holdout randomization test for feature selection in black box models, 2021. URL https://arxiv.org/abs/1811.00645.

Anderson M. Winkler, Matthew A. Webster, Diego Vidaurre, Thomas E. Nichols, and Stephen M. Smith. Multi-level block permutation. *NeuroImage*, 123:253–268, 2015. doi: 10.1016/j.neuroimage.2015.05.092.