IDENTIFIABILITY PRECEDES GENERALIZATION: A PRE-FIT AUDIT FOR TARGET–DOMAIN ALIASING

Anonymous authors
Paper under double-blind review

ABSTRACT

When the target label is determined by the acquisition domain, no amount of internal predictive performance establishes target-specific attribution from the observed data alone. Every cross-validated score computed from the labels equals the one computed from the domain-induced labels, so the competing explanations for the same number are indistinguishable. Distribution-shift evaluation asks how well a model trained on one distribution performs on another. This audit asks a different and prior question: what the recorded design allows a result to be attributed to. The two are complementary.

The audit has a pre-fit structural part and a model-based diagnostic part. Structurally, under an additive target–domain model the classical connectivity of the bipartite support graph identifies which target main-effect contrasts are estimable, including in designs that are incompletely crossed, where some class never appears in some domain. The singular values of the residualized target basis measure proximity to aliasing, which the rank alone cannot. Diagnostically, a classifier two-sample test conditioned within a target stratum tests for domain-associated differences the target model could exploit. We report its operating characteristics and the shifts it misses. For a *binary* target and a score carrying no target information beyond the domain, we give an exact identity, valid at any class prevalence, between that score's AUC against the target and its AUC against the domain, given the class–domain allocation.

On three datasets the audit separates different threats. On a sensor benchmark every additive target main-effect contrast is estimable despite incomplete crossing, yet random splitting reports accuracy 11.2 percentage points higher than leave-one-batch-out evaluation. In our Camelyon17 analysis the design is sound and hospital separability is nonetheless high: strong domain information and estimable target contrasts coexist. Grouping folds by slide lowers the reported AUC more than the subsequent hospital holdout does. In a clinical breath cohort disease status is inseparable from acquisition series, and two *healthy* series separate at $\mathrm{AUC} = 1.00$. The audit tells a study which contrasts its design supports and what its evaluation must block on. Applied to our own analysis it withdrew a date-prediction result, $r_s = 0.91$ under random folds and 0.48 once folds are grouped by collection date.

1 INTRODUCTION

A standard classification pipeline, applied to breath mass-spectrometry data from 127 healthy volunteers, separates them into two groups at $\mathrm{AUC} = 1.00$ under cross-validation. The two groups are not cases and controls. Everyone in both is healthy. They differ only in which acquisition series they were collected in.

Nothing is wrong with the pipeline. The features are real, the cross-validation is correctly stratified, no label leaked into a test fold. The score is a faithful answer to the question cross-validation asks—*can Y be predicted?*—and that question is not the one the study intended, which is *is Y being predicted for the intended reason?* When the target is a deterministic function of the acquisition domain, the two questions are not merely different: every cross-validated score, every model ranking and every calibration curve computed from (X, y) is numerically identical to the one computed from

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.

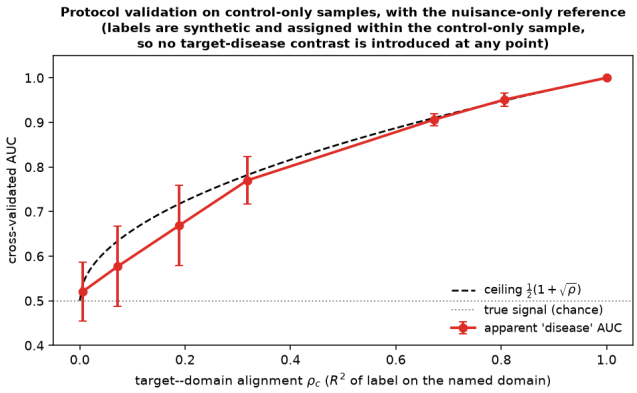


Figure 1: Apparent accuracy is a function of allocation, not of biology. Synthetic binary labels are assigned within a *control-only* sample of 127 participants, with probability p of matching acquisition-series membership; no target–disease contrast is introduced at any point on this curve. Cross-validated AUC rises from chance to 1.00 as alignment grows, and partial alignment already yields ≈0.77. The dashed line is the nuisance-only reference of Proposition 2, computed from a separately constructed domain score.

$(X, g(c))$ where c is the domain. No procedure operating on the observed data can separate them, and a perfect internal score establishes nothing attributable.

Two established lines of work sit close to this and answer different questions. Distribution-shift benchmarks measure how predictive performance transfers from one distribution to another (Koh et al., 2021; Gardner et al., 2023; Liu et al., 2023), and methods learn representations that transfer better. That question is well posed whatever the source design looks like: predicting the recorded labels is a well-defined task even when the target's contribution to the features cannot be separated from acquisition. *Well-defined* and *identifiable* are not the same property, and shift evaluation does not claim the second. Leakage audits and reporting checklists (Kapoor & Narayanan, 2023; Kapoor et al., 2024; Collins et al., 2024) likewise ask searching questions about process, splits and reporting—whether the estimate was computed correctly.

Our question is prior to both: given the design that was actually run, what may a result be attributed to? A study can pass every leakage check, evaluate correctly under shift, and still be unable to say whether its score reflects the target or the circumstances of collection. That is not a defect in either line of work; it is a question neither was built to answer, and it is the one an aliased design makes unavoidable.

We give the check. Our contributions are:

(1) A structural condition. Build the bipartite support graph on target classes and domain levels, with an edge wherever that cell holds an observation. Under an *additive* target–domain model its connectivity is exactly what decides estimability: k connected components leave $L - k$ independent target main-effect contrasts estimable. Full crossing is sufficient but *not* necessary—a design in which no single domain contains every class can still support every contrast, because additivity chains a contrast observed in one domain to one observed in another—and the condition is decidable from the design alone, before any model is fitted. The underlying result is classical (Searle, 1971; Rao, 1973); its use as an evaluation precondition is not.

(2) An exact identity for what confounding alone delivers. For a score whose information about the label flows entirely through the domain, we derive $\mathrm{AUC}(s, Y) = \frac{1}{2} + (a - b)(\kappa - \frac{1}{2})$ at arbitrary prevalence, with a, b the class-conditional domain probabilities and κ the domain separability of that same score, and generalize it to any finite number of domains. This is a *nuisance-only reference*: the performance a study would report from a score that reads acquisition and nothing else, and therefore the number a disease claim has to be argued against rather than compared favourably with.

(3) A diagnostic with measured operating characteristics. Same-target domain separability κ is a classifier two-sample test conditioned within a label stratum. We report its calibration (type-I error 0.030–0.045), its power (a Mahalanobis domain displacement $\delta \geq 1.5$ detected at n=80), and its blind spots—it misses covariance-only and mean-preserving mixture shifts, which a multi-bandwidth permutation MMD companion detects in our stress test. A diagnostic offered without operating characteristics is an opinion.

(4) Evidence that it changes conclusions. On three datasets the same procedure returns three different verdicts. A public sensor-drift benchmark is estimable despite incomplete support, and batch-agnostic evaluation is optimistic by 11.2 accuracy points. Camelyon17 is fully crossed with domain and target statistically independent, yet same-hospital separability is $\kappa = 0.969$ and the dominant evaluation error is the unit of analysis, not the domain. Our breath cohort is unidentifiable, and its apparent AUC can be tuned from 0.52 to 1.00 by allocation alone.

Applied to our own work, the framework withdrew a claim we had made: a date-prediction association that date-grouped validation reduces from $r_s = 0.91$ to 0.48. We report that alongside the rest, because a diagnostic that never costs its authors anything has not been tested.

2 Related work

Evaluating under distribution shift. A large literature measures how models transfer between distributions and builds benchmarks to do so (Koh et al., 2021; Gardner et al., 2023; Liu et al., 2023). It answers a question we do not: how a model behaves when the distribution changes. We ask what the source design permits a result to be attributed to, which is a separate property—in WILDS, tumour presence is defined independently of the hospital that produced the slide, so both questions are available there; in the cohort that motivates this paper only the first is. Our Camelyon17 analysis shows the two working together rather than in competition: the design is sound, the audit returns the split scheme the benchmark already recommends, and it adds a warning about the unit of analysis that the benchmark does not make.

Leakage, reporting standards and reproducibility. Recent work documents how often published pipelines leak information between splits and proposes checklists to prevent it (Kapoor & Narayanan, 2023; Kapoor et al., 2024), and clinical reporting standards now cover prediction models in detail (Collins et al., 2024; Moons et al., 2025). We adopt their discipline—all sample-coupled preprocessing is fitted inside training folds here—but the questions differ. A checklist asks whether the estimate was computed correctly. We ask whether the estimand exists. A study can satisfy every item of a leakage checklist and still report an AUC that no amount of transparency makes attributable, because the contrast it estimates is not identified by the design.

Batch effects and correction. Correction methods—ComBat (Johnson et al., 2007), WaveICA 2.0 (Deng et al., 2021), NormAE (Rong et al., 2020)—execute on any design. What requires crossing is the *disease-preserving interpretation* of their output; under collinearity no such certificate is computable, and confound regression can additionally leak label information (Hamdan et al., 2023; Chyzhyk et al., 2022).

Negative controls. The closest analogue to κ is the negative control outcome of epidemiology (Lipsitch et al., 2010; Shi et al., 2020): a contrast known *a priori* to carry no effect, used to expose confounding the primary analysis cannot see. Classifier two-sample tests (Lopez-Paz & Oquab, 2017) supply the machinery; conditioning within a label stratum is what turns one into a negative control, and we add measured operating characteristics and a link to the performance a confounded design will report.

Identifiability. The algebra we use is classical. Estimable functions in rank-deficient linear models and the connectivity condition for two-way designs are standard (Searle, 1971; Rao, 1973), and we say so rather than presenting them as new. What we add is their use as an evaluation precondition—computed before a model is fitted, reported alongside performance, and shown to change the conclusions of real analyses.

3 WHEN IS A TARGET CONTRAST IDENTIFIABLE?

Write $X \in \mathbb{R}^{n \times q}$ for the features, y for the target with L classes, and c for the acquisition domain with K levels—batch, site, session, instrument, or several of these jointly. Let Z be a full-column-rank nuisance design built from c (an intercept plus reference-coded indicators) and $M_Z = I - P_Z$ its residual projector. Everything below is decidable from (y, c) alone, before a model is fitted.

3.1 A MODEL-FREE IMPOSSIBILITY

Observation 1 (Non-identifiability)**.** *If $y_i = g(c_i)$ for every i, then for any pipeline the empirical label risk, every cross-validation score and every model ranking computed from (X, y) are numerically identical to those computed from $(X, g(c))$. Consequently no procedure using only (X, y) can distinguish a classifier exploiting target-related variation from one exploiting only domain-related variation.*

The proof is immediate—the two label vectors are equal entry-wise—and that is the point: the obstruction is arithmetic, not statistical. It assumes nothing about $P(X \mid Y, C)$, holds for any model class and any pipeline, and no amount of data, regularization or capacity removes it. Everything else in this section is weaker in that respect, and we mark where. It is *not* the claim that predicting y and predicting c are the same problem, which is false whenever $K > L$; it is the claim that predicting y and predicting $g(c)$ are, which is what forecloses attribution. The design also fails to identify the underlying parameters: in the linear mean model $\mathbb{E}[X \mid y, Z] = Z\Gamma + y\beta^\top$, collinearity places (β, Γ) on an affine subspace along which the fitted mean is constant and which contains $\beta = 0$, so a covariate-preserving correction can execute but nothing computable from the data certifies what it retained (Appendix A).

3.2 STRUCTURAL ESTIMABILITY UNDER AN ADDITIVE DESIGN

Non-identifiability is not all-or-nothing once $L > 2$, but saying which contrasts survive requires a model. We assume the additive form $\mathbb{E}[X \mid Y, C] = Z\Gamma + GB$, in which the target effect is the same in every domain, and ask which target main-effect contrasts it identifies. Build the bipartite *support graph*: one node per target class, one per domain level, an edge wherever that cell holds at least one observation. If it has k connected components containing class nodes, then

$$\Delta \;=\; \operatorname{rank}[Z, G] - \operatorname{rank}[Z] \;=\; \operatorname{rank}(M_Z G) \;=\; L - k$$

for any full-column-rank target design G, and the estimable target contrasts are exactly those supported *within* a component. Full Cartesian crossing gives $k = 1$ and is therefore sufficient for complete estimability—but it is not necessary. A design in which no single domain contains every class can still be connected, and then every contrast survives. The underlying result is classical (Searle, 1971; Rao, 1973); we verify $\Delta = L - k$ on six designs and use it as an evaluation precondition.

Additivity is doing real work here and we flag it rather than let it pass. Connectivity identifies a contrast between two classes that never share a domain by chaining contrasts that do—A against B in one domain, B against C in another—which is valid precisely because the additive model asserts that the A versus B difference is the same wherever it is measured. Under arbitrary $Y \times C$ interaction it is not: a cell with no observations leaves $P(X \mid Y{=}\ell, C{=}c)$ unidentified for any model, and no graph property recovers it. The impossibility of the previous subsection needs no such assumption; this positive result does, and a study relying on it is relying on additivity whether or not it says so.

This separates two things the literature runs together. *Support* is whether observations exist in a cell and gives the regime below; *estimability* is whether the requested contrast lies in $\operatorname{row}(M_Z G)$, and the two can disagree. Rank is also discontinuous in the design—it changes only when a cell goes from empty to occupied, and a design counts as aliased only when the target lies *exactly* in $\operatorname{col}(Z)$—so we additionally report the singular values of $M_Z Q_Y$ on the canonical orthonormal target basis, which are the sines of the principal angles between the target and nuisance subspaces and measure how *nearly* aliased a design is (Appendix A).

3.3 WHAT CONFOUNDING ALONE DELIVERS

Knowing a contrast is unidentifiable says what cannot be concluded. It does not say what number a confounded design will report, and that is what a reader needs in order to judge a published score. For a score carrying no target information beyond the domain, it is exact.

Proposition 2 (Nuisance-only reference)**.** *Let s be a score with $y \perp s \mid c$—all of its information about y flows through the domain.* (i) Binary domain, arbitrary prevalence. *With $c \in \{0,1\}$, write $a = \Pr(c{=}1 \mid y{=}1)$, $b = \Pr(c{=}1 \mid y{=}0)$ and $\kappa = \mathrm{AUC}(s,c)$. Then*

$$\mathrm{AUC}(s,y) \;=\; \tfrac{1}{2} + (a-b)\big(\kappa - \tfrac{1}{2}\big).$$

(ii) General domain. *With $c \in \{1,\dots,K\}$ and $A_{jk} = \Pr(s_j > s_k) + \frac{1}{2}\Pr(s_j = s_k)$ for independent draws from the domain-conditional score laws,*

$$\mathrm{AUC}(s,y) \;=\; \sum_{j,k} \Pr(c{=}j \mid y{=}1)\;\Pr(c{=}k \mid y{=}0)\,A_{jk}.$$

Neither form assumes balance; under balanced marginals (i) reduces to $\frac{1}{2} + \sqrt{\rho}\,(\kappa - \frac{1}{2})$, which is the special case rather than the statement. Two readings must be kept apart. This is an identity about *one specified score*, not a bound over pipelines: a score carrying genuine target information violates $y \perp s \mid c$ and the identity does not apply to it at all. And AUC does not decompose additively over information sources, so the difference between an observed AUC and this value is *not* an estimate of target signal. What the identity supplies is a reference—the performance attainable by reading acquisition alone—against which a claim must be argued, not from which it may be subtracted. Verified by simulation across 27 parameter settings and, on our own control-only pool, by constructing the nuisance score explicitly and reproducing the identity to a mean absolute error of 0.011.

3.4 A SAME-TARGET DOMAIN DIAGNOSTIC

Definition 1 (Same-target domain separability)**.** For a pair of domains c_a, c_b and a target stratum $y{=}\ell$, let $\kappa_\ell(c_a, c_b)$ be the cross-validated AUC of a classifier trained to predict domain membership from X, using only samples with $y = \ell$.

Conditioning within a label stratum is what makes it a negative control: under the null that the domains differ only by noise, κ sits at chance, and any excess is domain structure that a target classifier could exploit without touching the target. A large κ does not by itself invalidate a contrast—under a crossed design it mandates domain-blocked evaluation—and a κ at chance does not license random splitting, because the test is blind to shifts that preserve the mean.

3.5 REGIMES

Connectivity of the support graph sorts designs into the three states of Table 1. Defining them on the graph rather than on cell counts matters once $L > 2$: a design in which every domain holds two of four classes has no domain carrying every class, yet if the classes chain together it is connected and every main-effect contrast survives. A cell-counting definition has no category for that design; the graph does. The state is necessary but not sufficient: every conclusion it licenses is conditional on estimability against the *full* nuisance design rather than the one variable a study happens to call the domain. Where the requested contrast is not estimable, that fact is the output—reported as such, rather than replaced by a weaker contrast that is.

4 THREE DESIGNS, THREE VERDICTS

A diagnostic that returns the same answer everywhere is not a diagnostic. We run the procedure unchanged on three datasets—a public sensor benchmark, a histopathology benchmark, and the clinical cohort of §1—and it separates them (Table 2).

Regime	Support-graph state	What may be concluded
R1 *disconnected*	the support graph has $k > 1$ components	Contrasts spanning components are not estimable, and no target-specific attribution across them is available; target-preserving correction is unidentified. Contrasts *within* a component remain estimable.
R2 *connected, incomplete*	$k = 1$, but some cell is empty	Every additive target main-effect contrast is estimable—connectivity chains a contrast observed in one domain to one observed in another—but target $\times$ domain interactions are not identified where cells are empty, and splits must block on domain.
R3 *fully crossed*	every relevant cell is occupied	As R2, with interactions identified and precision available in every cell. Splits must still block on domain and on any dependence the deployment will face.

Table 1: States of the support graph with respect to the named domain, and what each licenses under the additive model. The three are exhaustive and mutually exclusive for any L, which a definition by cell counts is not: a design in which every domain holds two of four classes falls outside "some domain carries every class" and "no domain carries more than one" alike, yet is simply connected-and-incomplete here. Every entry in the third column is additionally conditional on estimability against the *full* nuisance design—a study can be fully crossed on batch while operator, also recorded, spans the target—and for $L > 2$ that condition is contrast-specific, not $\Delta > 0$.

Table 2: The same procedure, three designs, three verdicts. Δ counts estimable target contrasts out of $L - 1$; κ is same-target domain separability, macro over eligible domain pairs. The evaluation column reports the metric each dataset is normally scored on, under a random split and under a domain-blocked one. The breath cohort has no blocked split to report: with every acquisition series carrying a single class, holding out a domain removes a class.

Dataset	Domain	Regime	$\Delta/(L-1)$	κ	Random $\rightarrow$ blocked
UCI gas sensors (Vergara et al., 2012)	10 temporal batches	R2	5/5	0.997	$0.985 \rightarrow 0.873$ accuracy (-11.2 points)
Camelyon17 (Koh et al., 2021)	5 hospitals	R3	1/1	0.969	$0.895 \rightarrow 0.884$ AUC (-0.011; but -0.055 more from splitting patches rather than slides)
Breath MS (this work)	6 acquisition series	R1	0/1	1.000	no blocked split exists

A public benchmark that is estimable despite incomplete support. The UCI gas sensor array drift dataset (Vergara et al., 2012) holds 13 910 measurements of 6 gases across 10 temporal batches spanning three years, with batch membership recorded. Only 52 of the 60 gas $\times$ batch cells are populated at $n \geq 25$ and only 5 batches contain every gas, so support is incomplete and the positivity regime is R2. Estimability is a different question with a different answer: the support graph is connected, $k = 1$, so $\Delta = L - k = 5$ and every gas contrast survives. Reporting only the strict multiclass overlap $O_L = 0.473$ would have suggested a design in trouble; reporting only Δ would have hidden that three batches contain no toluene at all. Both belong in the report.

Same-gas batch separability is near-perfect, and concentration is the obvious confounder—a batch differs from another in what was measured as well as when. We remove it three ways. Restricting to *exactly* matched gas $\times$ concentration setpoints leaves 257 pairs with macro $\kappa = 0.997$ (minimum 0.942); tertile-stratified matching gives 330 pairs at 0.998; regressing concentration out continuously gives 0.980 (minimum 0.671, and we report the strata that do lose separation rather than summarising them away). The domain signal is not a concentration artefact. Ignoring it is expensive: six-way gas classification scores 0.985 accuracy under random splits and 0.873 under leave-one-batch-out,

an optimism of 11.2 accuracy points, 12.2 balanced-accuracy points and 16.8 macro-F_1 points—absolute differences, in a design that is otherwise sound.

A crossed design where the domain is not the problem. Camelyon17-WILDS (Koh et al., 2021; Bandi et al., 2018) predicts tumour presence in lymph-node patches across 5 hospitals. We sample 10 slides per hospital, cap the patches drawn from each slide, and group *every* split by slide, asserting slide-disjointness in each fold; the final sample is 10 526 patches from 50 slides—10 per hospital, each contributing patches of *both* classes, which is why the slide count is 50 rather than one set per cell. That overlap is what makes slide grouping consequential: a slide's tumour patches cannot be trained on while its normal patches are tested. The slide is the independence unit, and uncertainty is clustered there. Features are the penultimate activations of a frozen ImageNet ResNet-50 with no fine-tuning—we are not chasing histopathology accuracy, and an off-the-shelf representation makes the point better, since the diagnostics work on whatever features a study happens to have.

The design is the opposite of the clinical one. Every hospital carries both classes, so the regime is R3 with $\Delta = 1$, and our sampling makes target and domain nearly independent ($\rho_c = 0.0019$). Yet same-target hospital separability is macro $\kappa = 0.969$ over 20 pairs (median 0.997, minimum 0.772): a frozen ImageNet representation recovers which hospital produced a patch almost perfectly, within each tumour class. Near-zero alignment with κ this high is a combination no alignment statistic can express, and it is the reason κ is worth reporting separately.

The evaluation result needs three numbers rather than two, because two different leakages are present and they separate. A naive patch-level random split gives $\mathrm{AUC} = 0.950$; grouping folds by slide drops it to 0.895; holding out whole hospitals gives 0.884 (per-hospital 0.825–0.917, slide-clustered 95% CI $[0.825, 0.912]$). So $+0.055$ of the total optimism comes from the *unit of analysis* and only $+0.011$ from the *domain*. On this dataset the dominant error a study would make is splitting patches rather than slides, and a domain-blocked evaluation that still split patches within hospitals would inherit most of the inflation while appearing rigorous. We report the decomposition rather than the total.

A clinical cohort that supports nothing. The breath cohort of §1 is 183 samples and 2647 m/z features, in which every patient group is also a separate acquisition series. The support graph splits into $k = 2$ components—one holding the control classes, one the patient classes, with no edge between them—so $\Delta = 0$ and no target contrast is estimable. The regime is R1, and the framework's output is that fact rather than a performance number.

Two consequences are worth stating because they are what a reader of a published breath study would want to know. First, the same pipeline separates two *same-label* series of healthy controls at $\mathrm{AUC} = 1.00$, and $\kappa = 1.00$ on the control-only pool: the features carry acquisition information sufficient to saturate any binary contrast aligned with it. Second, on control-only samples with synthetic labels assigned at varying alignment to the acquisition series, apparent AUC moves from 0.52 to 1.00 by allocation alone, and partial alignment already yields ≈0.77—a value routinely published as a promising biomarker. No target–disease contrast exists anywhere on that curve.

5 Does the procedure track a design as it degrades?

A diagnostic that fires on our cohort and not on the benchmarks might be detecting something real, or might be detecting mass spectrometry. We test it on a design whose answer we control. Starting from the sensor benchmark, we degrade the crossing continuously: with probability p a sample's gas label is replaced by the one preferred by its batch, sweeping p from 0 (crossed) to 1 (aliased), over 100 independent randomized constructions per rung.

The procedure follows the design down. Alignment rises from $\rho_c = 0.012$ to 1.000 and overlap falls from $O = 0.924$ to 0.000; the regime moves from R3 to R1; and the conditioning spectrum—which rank cannot see—decays smoothly through $0.994 \to 0.959 \to 0.877 \to 0.732 \to 0.552$ before collapsing, while the optimism of ignoring the domain climbs from $+0.007$ to $+0.258$. Batch correction fitted on the training fold decays to chance over the same sweep, which is what Observation 1 predicts: at the aliased end the batch mean *is* the class mean.

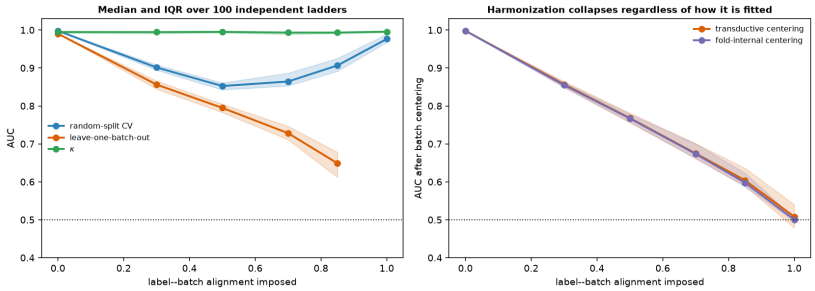


Figure 2: The procedure tracks a design as its crossing is destroyed. Median and interquartile range over 100 independent randomized constructions per rung. **Left:** as the label is made progressively a function of batch, the regime moves R3 $\rightarrow$ R1 and the optimism of batch-agnostic evaluation grows. **Right:** batch correction fitted on the training fold decays to chance over the same sweep, which is what Observation 1 requires—at the aliased end the batch mean is the class mean.

Two negative controls matter as much. Random partitions *within* a single acquisition series—where no domain contrast exists to find—return chance (medians 0.497 and 0.500 over 500 partitions each), so the diagnostic is not merely detecting that a partition was drawn. And κ has measured blind spots: it misses covariance-only and mean-preserving mixture shifts, which we detect with a multi-bandwidth permutation MMD companion in a stress test. We report the blind spots because a test offered without them invites the reading that a κ at chance licenses random splitting, which it does not.

6 Discussion

The framework answers a narrow question: given the recorded design, is the target contrast estimable, and if so how strongly. It does not tell a study which features are biologically real, nor does crossing license causal attribution—crossing makes the target contrast estimable with respect to the *measured* domain, and an unrecorded one can still confound it. Positivity on a named domain is likewise necessary and not sufficient: a design can be adequately crossed on batch while operator, also recorded, spans the target, which is why every regime verdict is conditional on estimability against the full nuisance design.

The most common objection to a diagnostic like this is that it is unfalsifiable—that it will always find something to complain about. Three things in this paper are offered against that reading. The procedure returns three different verdicts on three designs, including one where it says the design is fine and the real problem is elsewhere. It returns chance on constructions where the correct answer is chance. And applied to our own analysis, it withdrew a claim: a date-prediction association we had reported at $r_s = 0.91$ falls to 0.48 under date-grouped validation, because the original estimate let samples from one collection day appear on both sides of a split. We report that because a diagnostic that has never cost its authors anything has not been tested.

What we would ask of a study in an aliased design is not that it stop publishing. It is that it report the support matrix, the estimability of the contrast it claims, and a same-target domain control—and that where the contrast is not estimable, it say so rather than report a number that the design cannot support. The cost of doing this is a few lines of code. The cost of not doing it is a literature of AUCs that no amount of external validation can interpret, because the quantity they estimate was never defined.

Use of AI assistance

Per the ICLR 2027 policy we describe how large language models were used in preparing this work.

Writing. An LLM drafted portions of the prose, including first drafts of the abstract, introduction, theory section, experimental narrative and related work, working from results, decisions and argu-

ments supplied by the authors. Every draft was reviewed, edited and in most cases substantially rewritten by the authors.

Research ideation. Several framing decisions emerged from exchanges between the authors and the llm rather than from either alone. The support-graph formulation of estimability, the decision to test the framework on a second modality, and the separation of the model-free impossibility from the additive-design estimability result all originated this way. The underlying mathematics is classical and cited as such; the llm's role was in selecting and organizing it, not deriving it.

Code. Analysis scripts were written with LLM assistance and reviewed by the authors. All results reported in the paper are produced by those scripts and none is generated, estimated or reported by an LLM. The artifact includes a verification suite: claim checks, a numeric audit tying every reported figure to the analysis that produced it, figure-provenance records and unit tests on the theoretical machinery exist partly to make LLM-assisted analysis auditable.

Retrieval. LLM tools were partly used to locate and check related work. Every citation was verified against the primary source by the authors; none was accepted on the assistant's report alone.

What was not delegated. Among others, the research questions, the decision of what to claim and what to withdraw, the interpretation of every result and the final text are the authors'. No mathematical claim in this paper was proved by an LLM and no experimental result was produced by one.

REFERENCES

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015. doi: 10.1214/15-AOS1337.

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004.

Darya Chyzhyk, Gaël Varoquaux, Michael Milham, and Bertrand Thirion. How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience*, 11:giac014, 2022. doi: 10.1093/gigascience/giac014.

Gary S. Collins, Karel G. M. Moons, Paula Dhiman, Richard D. Riley, Andrew L. Beam, Ben Van Calster, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. doi: 10.1136/bmj-2023-078378.

Kui Deng, Fanjing Zhao, Zhiwei Rong, Lei Cao, Liang Zhang, Kang Li, Yan Hou, and Zheng-Jiang Zhu. Waveica 2.0: a novel batch effect removal method for untargeted metabolomics data without using batch information. *Metabolomics*, 17(10), 2021. doi: 10.1007/s11306-021-01839-7.

Josh Gardner, Zoran Popović, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with TableShift. In *Advances in Neural Information Processing Systems 36 (NeurIPS Datasets and Benchmarks Track)*, 2023. URL https://openreview.net/forum?id=XYxNklOMMX.

Sami Hamdan, Bradley C Love, Georg G von Polier, Susanne Weis, Holger Schwender, Simon B Eickhoff, and Kaustubh R Patil. Confound-leakage: confound removal in machine learning leads to leakage. *GigaScience*, 12:giad071, 2023. doi: 10.1093/gigascience/giad071.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007. doi: 10.1093/biostatistics/kxj037.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Sayash Kapoor, Emily M. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica Hullman, Michael A. Lones, Momin M. Malik, Priyanka Nanayakkara, Russell A. Poldrack, Inioluwa Deborah Raji, Michael Roberts, Matthew J. Salganik, Marta Serra-Garcia, Brandon M. Stewart, Gilles Vandewiele, and Arvind Narayanan. REFORMS: Consensus-based recommendations for machine-learning-based science. *Science Advances*, 10(18):eadk3452, 2024. doi: 10.1126/sciadv.adk3452.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010. doi: 10.1097/EDE.0b013e3181d61eeb.

Jiashuo Liu, Tianyu Wang, Peng Cui, and Hongseok Namkoong. On the need for a language describing distribution shifts: Illustrations on tabular datasets. In *Advances in Neural Information Processing Systems 36 (NeurIPS Datasets and Benchmarks Track)*, 2023.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations (ICLR)*, 2017.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72(4):417–473, 2010. doi: 10.1111/j.1467-9868.2010.00740.x.

Karel G. M. Moons, Johanna A. A. Damen, Tabea Kaul, Lotty Hooft, Constanza Andaur Navarro, Paula Dhiman, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*, 388:e082505, 2025. doi: 10.1136/bmj-2024-082505.

C. Radhakrishna Rao. *Linear Statistical Inference and Its Applications*. John Wiley & Sons, New York, 2nd edition, 1973.

Zhiwei Rong, Qilong Tan, Lei Cao, Liang Zhang, Kui Deng, Yue Huang, Zheng-Jiang Zhu, Zhenzi Li, and Kang Li. Normae: Deep adversarial learning model to remove batch effects in liquid chromatography mass spectrometry-based metabolomics data. *Analytical Chemistry*, 92(7):5082–5090, 2020. doi: 10.1021/acs.analchem.9b05460.

Shayle R. Searle. *Linear Models*. John Wiley & Sons, New York, 1971.

Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. A selective review of negative control methods in epidemiology. *Current Epidemiology Reports*, 7(4):190–202, 2020. doi: 10.1007/s40471-020-00243-4.

Alexander Vergara, Shankar Vembu, Tuba Ayhan, Margaret A Ryan, Margie L Homer, and Ramón Huerta. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166–167:320–329, 2012. doi: 10.1016/j.snb.2012.01.074.

A ESTIMABLE CONTRASTS AND CONDITIONING

Which contrasts are estimable. With L classes, "is the target estimable?" is the wrong question: a design can be aliased for some contrasts and crossed for others, and a study is entitled to know which of *its* comparisons remain available. The answer is classical estimable-function geometry (Searle, 1971; Rao, 1973); our contribution is not the algebra but its operational use as an evaluation diagnostic under domain aliasing. Four questions separate cleanly.

Do observations exist? That is positivity, and it gives the regime—a statement about *support completeness*, and only that. We had been reading it as a statement about identifiability, and the external benchmark shows the two come apart: three of ten batches contain no toluene, so support is incomplete and the design is R2, yet all five gas contrasts are estimable. R2 is a support category.

The classical two-way theory says exactly why (Searle, 1971; Rao, 1973). Build the bipartite *support graph*: one node per target class, one per domain level, an edge wherever that cell holds an observation. If it has k connected components containing class nodes, then

$$\Delta \;=\; L - k,$$

and the estimable target contrasts are precisely those supported *within* a component. Full Cartesian crossing gives $k = 1$ and is therefore sufficient for complete estimability—but it is not necessary. We verify the identity on six designs (`support_graph.py`), including a chain in which *no* domain holds all four classes yet the graph is connected and every contrast survives: positivity calls that design R1, and it is fully estimable.

The graph also gives R1 a structural reading rather than an arithmetic one. Our breath cohort has $L = 2$ and $k = 2$: controls and patients occupy separate components with no edge between them, so $\Delta = 0$. The components name what can never be compared—{H, `CTRL-B`} against {KI, B, C, LU}—and equally what can. A *within*-component contrast like H versus `CTRL-B` remains estimable, which is exactly the negative control we run in §D: in a disconnected design, the estimable contrast is the one that carries no target information at all.

How many target directions survive? Write $M_Z = I - P_Z$ for the residual projector of the nuisance design and let G be *any* full-column-rank target design—reference coding will do. Then

$$\operatorname{rank}\big(M_Z G\big) \;=\; \operatorname{rank}[Z, G] - \operatorname{rank}[Z] \;=\; \Delta,$$

so the multiclass count and the rank gap of §3 are one quantity written two ways, and we use Δ for both. Rank is invariant to any nonsingular recoding of G, so the coding genuinely does not matter here—a point we return to below, because it is *not* true of the conditioning spectrum. $\Delta = 0$ means the graph splits into one component per class, so no target contrast survives ($\operatorname{col}(G) \subseteq \operatorname{col}(Z)$, the R1 condition); $\Delta = L - 1$ means the graph is connected—which full crossing implies but does not require; and $0 < \Delta < L - 1$ is *partial identifiability*, a case with no binary analogue. On a constructed design in which two of four classes are crossed across domains and two are confined to their own we recover $\Delta = 1$ of 3.

Which contrasts, specifically? A target contrast is a weight vector h over the classes summing to zero—A versus B is $h = (1, -1, 0, 0)$—with coordinates $a = S^{-1}V^{\top}h$ in the orthonormal contrast basis, where $Y^{\star}SV^{\top}$ is the singular value decomposition of the centred class-indicator matrix. The estimable set is the *row space* $\mathcal{E} = \operatorname{row}(M_Z Y^{\star})$, and $a^{\top}B$ is estimable exactly when $a \in \mathcal{E}$, equivalently $a \perp \operatorname{null}(M_Z Y^{\star})$. It is *not* enough that $M_Z Y^{\star} a \neq 0$: with $M_Z Y^{\star} = [\,1 \;\; 0\,]$ and $a = (1, 1)^{\top}$ that product is nonzero, yet a has a component along a wholly unidentified direction. Where $a \notin \mathcal{E}$, only $P_{\mathcal{E}} a$ is identified, and we summarise the position by the *alignment*

$$\eta(a) \;=\; \frac{\|P_{\mathcal{E}}\, a\|}{\|a\|} \in [0, 1],$$

which is 1 when the contrast lies entirely in the estimable subspace and 0 when it lies entirely in the complement. We call it alignment rather than "fraction identifiable" because it is a projection, and the latter invites a stronger statistical reading than the geometry supports. On the design above, A versus B gives $\eta = 1.000$, C versus D gives 0.000, and A versus C gives 0.513—mixing an estimable direction with an aliased one. A study asking whether its own comparison survives gets an answer about that comparison.

How strongly? Rank is a discontinuous function of the design: it changes only when a cell goes from empty to occupied, and a design counts as aliased only when the target lies *exactly* in $\mathrm{col}(Z)$. One observation therefore flips the verdict while leaving the contrast just as weakly supported. Conditioning is the continuous answer, but it requires more care than rank does: singular values are *not* invariant to recoding. Multiplying one column of G by 100 leaves Δ unchanged and changes the spectrum entirely, so a conditioning diagnostic must be defined on a canonical basis.

We use the orthonormal target basis. Let H be the $n \times L$ class-indicator matrix, centre it as $H_c = M_{\mathbf{1}} H$, and take $H_c = Q_Y R$ with $Q_Y^\top Q_Y = I_{L-1}$. The conditioning spectrum is

$$\sigma_1 \geq \cdots \geq \sigma_{L-1} \;=\; \mathrm{svd}\big(M_Z Q_Y\big),$$

which is coding-invariant, lies in $[0, 1]$, and has a geometric reading: the σ_i are the *sines of the principal angles* between the target subspace and $\mathrm{col}(Z)$. A direction orthogonal to the nuisance space contributes $\sigma = 1$; one lying inside it contributes $\sigma = 0$.

The binding quantity is $\sigma_{\min}$, not $\sigma_{\max}$, because the worst-supported direction determines the worst-conditioned contrast. On a design where two of four classes are confined to their own domains the spectrum is $(0.9769, 0, 0)$: the largest suggests the target is well separated from Z, and only the smallest reveals that a direction is exactly aliased. For a binary target there is one singular value, $\sigma = \sqrt{1 - R^2_Z}$ with R^2_Z taken against the *full* nuisance design; this equals $\sqrt{1 - \rho_c}$ only when Z is the named domain plus intercept, as it is in our ladder.

Structural estimability and conditioning are therefore different reports, and both belong in the template. Our ladder makes the gap visible: the design is R3 by positivity at every rung short of exact aliasing, while σ falls $0.994 \to 0.959 \to 0.877 \to 0.732 \to 0.552$ and the optimism of ignoring the domain climbs from $+0.007$ to $+0.258$. A single observation in an empty cell would flip the regime verdict; it would barely move σ, which is the quantity tracking how much the design actually supports. Rank answers whether a target direction is structurally identifiable; the principal-angle spectrum answers how close the design is to losing it.

B SUPPORT VERSUS ESTIMABILITY

Two different questions. The regime is decided by positivity with respect to one *named* domain; the parameter-aliasing result of §3 is stated for the whole nuisance design Z, which may carry site, instrument, operator, period and batch together. These are not the same question and they can disagree. Suppose every acquisition batch contains cases and controls—positivity holds, and the protocol as described so far calls the design R3—but a second measured axis, say operator, is perfectly aligned with the target. Then $\Delta_Z = 0$ on $Z = [\mathbf{1}, \text{batch}, \text{operator}]$ and no target contrast is estimable, even though the regime looks sound. We construct exactly this case (`nuisance_scope.py`): regime R3 on batch, $\Delta_Z = 1$ on batch alone, $\Delta_Z = 0$ on batch and operator jointly.

Both must therefore be reported, because they answer different questions. *Structural support* with respect to the named domain—R1, R2 or R3—says whether that domain's levels carry the contrast. *Algebraic estimability* Δ_Z, conditional on every nuisance variable measured, says whether anything recorded already spans the target. A design is safe only if it passes both. Note also that adding variables to Z can only reduce Δ_Z, never restore it: a study that measures more of its own provenance can discover that its contrast was never estimable, but never the reverse. That asymmetry is an argument for recording run order, operator and kit, not a reason to avoid doing so.

We leave the AUC identity of Proposition 2 explicitly binary-target: it concerns a binary decision problem, and nothing in the protocol requires a multiclass analogue.

C HOW MUCH DATA DOES THE TEST NEED?

This appendix characterizes κ against the domain displacement δ, the quantity that governs it within a label stratum. The test holds its nominal level (type-I error 0.030–0.045 for $n \in \{20, 40, 80\}$) and detects $\delta \geq 1.5$ at 80% power with n=80. The asymmetry matters: a *significant* same-target domain test is interpretable under the calibrated settings of §C, whereas an observed $\kappa > \frac{1}{2}$ on its own is not—its finite-sample null sits above chance—and a nonsignificant result on a small stratum is weak evidence of absence, and must carry the power achieved at that size.

D CONTROLS WITH KNOWN ANSWERS

A diagnostic that fires on every input is worthless, so we run it on contrasts whose answer is known in advance.

The falsification test is a random partition *within* a single healthy series: same participants, same collection period, an arbitrary split. If the pipeline separated those, nothing else in this paper would mean anything. It does not— median $\mathrm{AUC} = 0.497$ within $\texttt{CTRL-B}$ and 0.500 within H over 500 random partitions each—while the identifier-derived contrast between the same two series reaches 1.00. We report the distribution rather than the mean (the hyperparameter sweep) because the spread is what makes the control informative: individual partitions reach 0.66, so this test demonstrably can produce high values by chance, and the observed 1.00 lies far outside that range rather than merely above a flat baseline. The separation is a property of the series boundary, not of the pipeline's willingness to separate anything.

On the external benchmark the same three-way pattern holds. Within one (gas, concentration-tertile, batch) cell a random partition returns 0.506 (chance). Holding gas and concentration fixed but crossing batches returns 1.00—drift. Holding batch fixed but changing the gas also returns 1.00—legitimate class information. The pipeline is therefore silent where nothing differs, and equally loud for a nuisance difference and a real one, which is exactly why the *design* must decide which it has found: no statistic computed from the features alone can tell drift from disease.

One suggested control cannot be built. An exact- or nearest-date matched comparison of the two control series does not exist: they share no collection day, and the closest attainable pairing is 162 days apart. We report the absence rather than substituting a weaker contrast.

Relation to distribution-shift benchmarking. A parallel literature studies how models behave *under* distribution shift. TableShift (Gardner et al., 2023) assembles 15 real tabular tasks with explicit shifts, several clinical, and measures how far performance falls out of distribution; WhyShift (Liu et al., 2023) argues that observing degradation is not enough and that one must characterise *which kind* of shift occurred, finding $Y \mid X$ shifts to be the most prevalent in practice. Our question is logically prior to both. They presuppose a target contrast that is well defined in the source domain and ask how it transfers; we ask whether that contrast is separable from the domain structure in the first place. Under R1 there is nothing to transfer—the source-domain quantity is not identified—so a shift benchmark applied to such a design would faithfully measure the degradation of a number that never meant what it appeared to. Identifiability is a precondition for distribution-shift evaluation, not a special case of it.

E REPORTING TEMPLATE

The reporting template below gives the checklist in two layers: a *core identifiability layer* that applies to any modality, and *spectral and biomarker extensions* that apply where the feature space is low-dimensional or chemically interpretable. The split matters for scope: an imaging study should report its domain definition, alignment, regime and same-target control, but per-feature univariate performance over a million pixels would tell no one anything. It is meant to *complement*, not replace, the general reporting standards now available: REFORMS for machine-learning-based science (Kapoor et al., 2024), and TRIPOD+AI and PROBAST+AI for clinical prediction models (Collins et al., 2024; Moons et al., 2025). Those instruments ask, correctly, for transparency about data, splits, leakage and performance. What they do not ask—and what a study in an aliased design most needs to state—is whether the target contrast is *estimable at all*. That question has a hierarchy, and we state it in that order: **support**—the domain-by-class matrix, the overlap O, and the regime; **estimability**—the rank gap Δ and the row space $\mathrm{row}(M_Z Y^\star)$ relative to the *full* nuisance design; **conditioning**—the singular values of $M_Z Y^\star$, which say how nearly aliased a design is where rank only says whether it is; and **separability**—κ with a nonlinear or kernel companion. Evaluation follows from those, not the reverse. A study can satisfy every item of a general checklist and still report an AUC that no amount of transparency makes attributable. Our core items are the identifiability layer those checklists sit on top of.

F METHODS

F.1 PREPROCESSING

Untargeted HRMS (Study A). Raw intensities (183×2647) contain no missing values and span nine orders of magnitude. Each spectrum is total-ion-current (TIC) normalized to a *fixed* reference (10^{11}), then $\log_{10}(x+1)$ transformed. The reference is deliberately a constant rather than the dataset median TIC: the median would be estimated from all samples, and $\log(x+1)$ is not equivariant to a multiplicative constant, so scaling to it would not be strictly row-wise. With a fixed reference each sample is transformed using only its own values, so no information flows between samples and the step may be applied outside cross-validation; the choice of reference is immaterial to our conclusions (the leakage-cost analysis). Feature standardization is *not* applied here at all; see §F.2. **Targeted panel.** The 6 calibrated concentrations (ppb) are $\log_{10}$-transformed after clipping at 10^{-3} ppb to bound zeros. The clip is inert here: the panel contains no zeros and no missing values, and the global minimum across all six analytes is 5.7 ppb, so no value is affected. Reported concentrations are the provider's calibrated outputs; per-analyte limits of detection and quantification were not supplied with the export . **Study B.** Positive-mode spectra are filtered to features detected (> 0) in $\geq$30% of samples ($8215 \rightarrow 7531$), then $\log_{10}(x+1)$ transformed. **Laboratory fields.** One serum creatinine value (record KI-14, 83.0 mg/dL) is physiologically implausible and internally inconsistent: its paired urea (109) is elevated but unremarkable in this cohort, whereas that creatinine would indicate near-terminal renal failure; a misplaced decimal (8.3 mg/dL) fits both. We exclude this value and retain all others, including 27 creatinines between 5.5 and 13.6 mg/dL that are plausible in kidney insufficiency. The exclusion is implemented as a pre-specifiable rule—any serum creatinine above 20 mg/dL is set to missing—rather than as a judgement about one record; on these data the rule fires exactly once. Rank-based statistics are used throughout, but they guard against outliers, not against unit or transcription errors .

F.2 CROSS-VALIDATION AND PREPROCESSING DISCIPLINE

Because this paper concerns evaluation validity, we hold our own pipeline to the standard we advocate. *All* sample-coupled preprocessing—the Study B prevalence filter, feature standardization, and PCA where used for prediction—is fitted **inside each training fold** via a `scikit-learn Pipeline`, never on the full dataset, avoiding the preprocessing leakage documented by Kapoor & Narayanan (2023). PCA appearing in figures and in variance partitioning is descriptive only and is explicitly marked as such. Study A uses stratified 5-fold CV. Because the mean of fold-wise AUCs is not the same estimator as the pooled out-of-fold AUC—and can behave poorly when folds hold few positives—we report both, together with their spread over 20 complete CV repetitions rather than a single seed. For the targeted panel we report the 2.5/97.5 percentiles of the 100 fold scores from 20×5-fold repeated CV. We label this a *split-sensitivity interval* and not a confidence interval: fold scores share training data and reuse observations, so their percentiles do not estimate sampling variability, and no unbiased estimator of the variance of K-fold CV exists (Bengio & Grandvalet, 2004). Sampling uncertainty is reported separately as a stratified participant-level bootstrap over fixed cross-fitted out-of-fold predictions; for Study B the bootstrap and the permutation null are both clustered by participant. An unconditional bootstrap that refits the entire pipeline per replicate is not well defined for this estimand: resampling participants with replacement either duplicates identities—making a 10-way identification task artificially harder, since one person appears as two inherently confusable classes—or deduplicates them, leaving fewer classes and making it easier. Both change the task rather than the sample, so we report the conditional interval and label it as such. Study B uses leave-one-(subject, day)-out CV, so an entire acquisition day is held out and a subject is recognized only from *other* days; the temporal analysis instead trains on each subject's earlier days and tests on their latest day. Every reported metric is accompanied by a permutation null: labels are shuffled and the identical procedure rerun B times, and we report the standard estimator $p = (1+\#\{T_b \geq T_{\text{obs}}\})/(B+1)$. Headline nulls use $B \geq 999$—B=1000 for the Study A tests and B=999 for the Study B block null and the external benchmark, giving smallest attainable p-values of 0.001 in both cases—with the exception of TabPFN, which uses B=200 (smallest attainable 0.005) because its inference cost far exceeds the baselines'. Permutations respect the dependence structure of each design. Study A permutes sample labels. One point about the *conditional* null generalizes beyond our data and is worth stating, since the protocol is offered for reuse: under R2 it must be computed on the supported domains alone. A domain carrying one class contributes la-

bels that a within-domain permutation cannot move, so retaining it leaves part of the target–domain association intact in the null and the test becomes anticonservative. On constructed data with *no* target effect whatever—so that a correct conditional null must sit at chance—permuting within all domains leaves the null centred at 0.519, whereas restricting to the supported domains gives 0.489, an inflation of 0.030 (`conditional_null_estimand.py`). The contrasts carried by unsupported domains are reported as non-estimable rather than allowed to contribute fixed labels. This does not affect our own results—Study A is R1, where the conditional null is empty rather than merely restricted—but it changes what the protocol asks of an R2 study. For Study B the estimand is cross-day identity consistency, so the null must destroy exactly that while preserving everything else. Let $\mathcal{B} = \{(s, d)\}$ be the set of $(\text{subject}, \text{day})$ blocks and $\sigma(\cdot)$ the subject owning a block. The *block-level* null draws a permutation π of the multiset $\{\sigma(b) : b \in \mathcal{B}\}$ and relabels every sample in block b with $\pi(b)$: all measurements taken on one day stay together and keep their within-day dependence, but a subject's days no longer share an identity. The exchangeability assumption is thus at the block, not the sample, level—days are exchangeable across subjects under the null, samples within a day are not. This is the null for the *subject-day* estimand, evaluated at the same budget as every other headline test, $B = 999$: the null mean is 0.084 against chance 0.100, its maximum over 999 draws is 0.207, and the observed 0.707 therefore exceeds every permutation, giving $p = 0.001$, the smallest attainable at this B. (A row-level shuffle, which ignores the within-day dependence that motivates the block null, gives a similar null mean of 0.095 over 30 draws; we report it only as a sanity check and draw no inference from it.) A *bijective* relabeling of subjects is by contrast *not* a valid null at all: identification is invariant to renaming, so it reproduces the observed statistic exactly and returns $p = 1$ regardless of signal.

Study B has *two* estimands and each requires its own null; pairing them wrongly is not a matter of taste. The **subject-day holdout**—a held-out day matched using that subject's other days—takes the block-level null above. The **whole-session holdout**—an entire calendar date removed, so that no sample from the test session appears anywhere in training—takes a *within-session identity permutation*, which permutes identities among the subjects actually present in each session. That preserves how many people attended, which people they were, every session-specific effect and the attendance structure, while destroying the cross-session correspondence of identity the estimand concerns. Applying the block-level null to the session-level holdout is what we tried first and it fails: with 9 sessions and 10 subjects, a block permutation leaves the assigned identity absent from training in every draw, so those samples are unpredictable by construction and the null falls *below* chance.

We quantify the cost of this discipline in the leakage-cost analysis.

F.3 Models and hyperparameters

Classifiers are deliberately simple, since the paper's claims concern evaluation rather than model capacity. ℓ_2-regularized logistic regression (C=0.1 for the 2647-dimensional HRMS features, C=1.0 for the 6-dimensional targeted panel and Study B principal components; `lbfgs`, `max_iter`=3000); random forest (400 trees, default depth); gradient boosting (scikit-learn defaults); and TabPFN v2 (Hollmann et al., 2025), a transformer pre-trained on synthetic tabular priors and suited to $n \lesssim 10^4$, $p \lesssim 500$—hence applicable to the calibrated VOC panel but not to raw HRMS without reduction. TabPFN uses B=200 rather than B=1000 permutations, as its inference cost far exceeds the baselines'. Study B uses $\min(40, n_{\text{train}} - 1)$ principal components before logistic regression. No hyperparameter search was performed. We say the values were *fixed without tuning on evaluation performance* rather than "fixed a priori": they are conventional defaults, chosen once and never adjusted in response to a score—which is what matters for selection bias—but they were not registered before the analysis began. the hyperparameter sweep varies each of them and finds the conclusions unchanged.

F.4 Statistical procedures

Group differences in calibrated concentrations use two-sided Mann–Whitney U with Benjamini–Hochberg FDR control across the 6 analytes; effect sizes are $\log_2$ fold-changes of medians. Within-subject diurnal contrasts use the Wilcoxon signed-rank test on subject means (one pair per subject, 7531 features, BH-corrected). Breath–laboratory associations use Spearman r_s on a pre-specified set: ammonia only—chosen on mechanism, urea being its nitrogen source and creatinine the stan-

dard renal marker—against each of the two laboratory values, cross-cohort and within the patient cohort, giving four tests, BH-corrected across that set. Robust feature sets use a bootstrap ℓ_1 *selection-frequency* analysis: 200 resamples with replacement, `liblinear`, a single fixed C=0.3, standardization refitted within each resample, and the per-feature selection frequency reported. We deliberately do not call this stability selection (Meinshausen & Bühlmann, 2010): that framework subsamples without replacement at $\lfloor n/2 \rfloor$ across a regularization path and derives an error bound from a selection threshold, none of which we do. The frequencies below are descriptive and carry no error control. Model-X knockoffs (Barber & Candès, 2015) would be the natural way to obtain it, and we note only why they are unavailable here: the guarantee requires the joint law of X to be known or well approximated, which is not credible for 2647 strongly correlated features at n=183.

F.5 Confound diagnostics

Cohort-explained variance is the between-group sum of squares over the total sum of squares, computed per feature block and aggregated (a one-way analogue of PVCA). Label–cohort alignment ρ_c is the R^2 of the label regressed on indicators for the *named* domain c. We do not compute it against the full nuisance design Z: R^2 rises mechanically as Z gains columns, so a study that recorded site, instrument, operator and period would report a larger alignment than one that recorded only batch, for reasons having nothing to do with confounding. ρ_c is descriptive and comparable only within a fixed choice of c; the algebraic questions—Δ, the row space, the conditioning—are the ones asked of the full Z, and they do not degrade in this way. Numerical rank needs an explicit tolerance, because the protocol is offered for arbitrary recorded nuisance variables rather than only 0/1 design matrices. We compute the singular spectrum, apply a relative tolerance of 1×10^{-9} *together with an absolute floor of the same size*, and record the spectrum alongside the verdict. The floor is not decoration: NumPy's default tolerance is relative to the largest singular value, so a numerically zero residual—exactly what an aliased design produces—is reported as *full* rank, inverting the verdict. We also report rank across tolerances spanning 1×10^{-12} to 1×10^{-4}, so a reader can see the answer is not an artefact of where the threshold was placed. We run this on every design in the paper—both real datasets and the four constructed ones—and Δ is invariant across the whole range in each (`support_graph.py`). Exact rank deficiency and near-singularity are then distinct findings, the latter reported through the conditioning diagnostic of §A. The aliasing test compares $\mathrm{rank}[\mathbf{1}, C]$ with $\mathrm{rank}[\mathbf{1}, C, y]$ under a stated numerical policy (with C the full one-hot coding, $[\mathbf{1}, C]$ is rank-deficient by construction; the test compares nested designs, so only the rank *difference* matters and the deficiency cancels); equality means the disease covariate is aliased with batch and cannot be preserved by covariate-aware harmonization. Normalization sensitivity re-runs the negative-control test under TIC, PQN, and no normalization.

F.6 Study B repeatability

Per-feature repeatability is the one-way (ICC(1)) intraclass correlation, $\mathrm{ICC} = (\mathrm{MS}_B - \mathrm{MS}_W)/(\mathrm{MS}_B + (n_0 - 1)\mathrm{MS}_W)$ where, for unequal group sizes, $n_0 = \big(N - \sum_i n_i^2/N\big)/(k-1)$ rather than the arithmetic mean of the n_i; with our group sizes the two differ by under 2%, shifting the median ICC by <0.005. A random-intercept mixed model would be cleaner still for irregular designs, but fitting one per feature across 7531 features to move the estimate by that margin is not a use of effort we can justify, and we say so rather than implying the choice is immaterial in general. Values are reported unclipped; negative values (within-subject variance exceeding between-subject) are retained in all summary statistics and truncated only for histogram display.

F.7 Harmonization and its limits

We consider ComBat (Johnson et al., 2007), WaveICA 2.0 (Deng et al., 2021), and NormAE (Rong et al., 2020). Covariate-preserving harmonization regresses out batch while retaining a biological covariate, which requires the covariate to vary *within* batches. Under collinearity (y=$g(c)$) it cannot be specified without removing the label. The obstacle is identification, not execution: each method runs perfectly well here—ComBat with an empty covariate matrix, the unsupervised correctors with no batch labels at all—and each returns transformed data. What cannot be done is certify what survived. We therefore report that *disease-preserving harmonization is unidentified* rather than that

harmonization is impossible, and establish it via the rank test above; naive confound regression additionally risks leakage (Hamdan et al., 2023).

F.8 Reproducibility and compute

One command, `make verify`, runs every analysis, re-derives each reported claim from the data (`verify_claims.py`), runs the theory tests (`test_theory.py`), and audits the manuscript on three surfaces: numbers (`audit_numbers.py`), figure provenance (`figure_provenance.py`), and retired claims and internal labels (`lint_manuscript.py`). For the numeric audit, every `\num{}` in the source is matched against the values the analyses actually wrote, and any number with no such support is listed. That audit exists because the claim checker cannot see values that were typed into the text or drawn into a figure; it found 27 such values on its first run, including six supporting a withdrawn claim that no script in the artifact then reproduced. All are now either produced by a script or annotated with their source. We state its limit as plainly: the audit establishes that the manuscript agrees with the results store, *not* that the store is correct—that is what the claim checks and the unit tests are for. The latter (`test_theory.py`, 32 assertions) exercise the theoretical machinery on constructed inputs whose answers are known in advance: the regime classifier under binary and multiclass targets, complete aliasing, partial support and relabelling of both targets and domains; the the parameter-aliasing result of §3 perturbation, including the point at which the target effect vanishes and a crossed control where the same perturbation must *fail*; the Proposition 2 identities at arbitrary prevalences and for K domains, together with a case in which the balanced special case and the general form must diverge; and the numerical rank policy, including a matrix on which NumPy's default relative tolerance returns *full* rank where the correct answer is zero. All analyses are CPU-only and complete in a few hours on a standard workstation (8 cores, 30 GB RAM); no GPU is required.

We are precise about what an external reader can reproduce. The breath and clinical data are governed and cannot be released, so the reported *numbers* are not independently reproducible outside the consortium. The artifact therefore also ships a seeded generator for a structurally faithful *synthetic* dataset—same file layouts, identifier conventions, cohort sizes and date structure, with a cohort effect and label–cohort aliasing deliberately injected—on which the complete workflow executes unmodified (`make verify DATA=data_synthetic`) and reaches the same qualitative verdicts. A reader can therefore run and inspect every step of the analysis without access to clinical data, while the quantitative results in this paper rest on the governed data. Random seeds are fixed for all splits, permutations, and bootstraps, and each analysis draws from an *independent* seed-derived generator rather than a shared global stream—so results do not depend on execution order or on which optional packages (e.g. TabPFN) are installed. Exact library versions and wall-clock runtime are captured automatically into the released `results.json` at run time; the reference environment is Python 3.11 with `scikit-learn`, `numpy`, `pandas`, `scipy`, and `matplotlib` pinned in the accompanying `environment.yml`.