MODALITY VALUE IS A PAIRED COMPARISON: CENSORING- AND ACQUISITION-AWARE ESTIMATION OF OMICS MODALITY GAIN

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal survival research typically measures an omics modality's value by the performance gain from adding it, a metric that shapes model design, modality selection, and testing resource allocation. In practice, however, omics testing covers only a subset of patients, its availability can depend on patient characteristics, disease status, and clinical workflow, and survival outcomes are further subject to right censoring—making the modality's true gain in the target population hard to obtain from the observed subset. Existing weighting methods typically estimate performance with and without the modality separately, turning a paired target into a difference of separately normalized estimates, which can inflate variance and turn "not estimable" into "no value." We define modality value as the paired performance difference of a fixed predictor under controlled-input ablation, derive identification and non-identifiability conditions, and propose a model-agnostic framework: joint weighting corrects for selective acquisition and censoring, paired increments preserve score correspondence, support diagnostics flag unreliable cases, and rank discordance guides adaptive selection between estimators—without retraining or imputation. Across 10 TCGA cohorts, 3 model classes, 5 acquisition mechanisms, SurvBoard's natural missingness, and 10 cross-cohort transfers, naive estimation shows substantially higher error and flips the sign of the estimated gain in 26.0% of signal-present cases. A selection rule learned on semi-synthetic cohorts and frozen before external transfer improves over either fixed estimator.

1 INTRODUCTION

Research on multi-omics survival prediction often focuses on stronger fusion: multimodal graphs, attention, and imputation-or-generation architectures for missing modalities [27; 29; 30], while benchmarks such as SurvBoard [21] report per-modality contribution columns that guide model design and testing decisions. Both rely on an important premise: *the modality gain computed on observed data reflects the modality's value for the target population*.

That gain is a difference of two discrimination measures on the same patients, one from a score that uses the omics block and one from a score that does not. Such comparisons are naturally *paired*, because the two estimates share patient-level noise that can cancel in the difference [1]. Kang et al. [2] extended this idea to censored concordance indices. However, both approaches assume that every score is available for every subject. That assumption fails here: the full-input omics score exists only for acquired patients, while the omics-ablated score can be computed for everyone. Acquisition can also be selective—whether an assay is run can depend on payer, pathway, institution, and patient status, so missingness can be related to outcome ("informative presence"; 5). Right censoring adds another challenge. For some patients, the event time is not fully observed, so the patient pairs used to measure survival discrimination are also only partly observed. Modality value must therefore account for both selective acquisition and censoring while preserving the paired comparison. ICYM2I [3] showed an "illusion of informativeness" under selective observation for classification, but its framework does not address censoring or a survival estimand.

We take an estimand-first route and build no fusion or imputation module. We show that standard evaluation and natural IPW/IPCW corrections can fail for three distinct reasons: biased censoring

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.

estimation, unpaired differencing, and loss of effective support. The proposed estimators operate only at evaluation time and require neither retraining nor imputation.

Contributions.

1. **Structural problems in multimodal valuation (Three Defects D1–D3)** (Section 4). We show that modality gain under selective acquisition and censoring is a paired-comparison problem. We identify three defects in current evaluation: **D1**, censoring is estimated on the acquisition-selected subsample; **D2**, the paired target is computed as a difference of separately normalised estimates; and **D3**, effective support can be empty, yet pipelines still report a value. Naive gain underperforms reporting zero (0.103 vs. 0.081 error) and flips sign in 26.0% of the signal-present stratum. Standard corrections perform even worse.
2. **Paired incremental contrast and pair-kernel noise reduction** (Section 3, Appendix A.5). We express modality value as a paired increment rather than a difference of separately estimated values: $\Delta_m = E[\kappa(h^+ - h^-)]/E[\kappa]$. Directly weighting $d = h^+ - h^-$ removes acquisition-weight noise on pairs whose ordering does not change. Proposition 2 links this pair-kernel effect to the discordance probability $q = P(h^+ \neq h^- \mid \kappa = 1)$. We test the resulting empirical prediction without claiming an exact variance formula for the full U-statistic.
3. **A theory-motivated, empirically calibrated rule for when pairing helps** (Sections 3, 6.5). Pair-kernel analysis predicts that the advantage of paired weighting decreases with q, which we validate in semi-synthetic and external-transfer settings. A semi-synthetic boundary, $P_{\text{eff}} = 0.306\,[0.258, 0.374]$, is frozen before transfer and yields mean absolute error 0.034 on 20 transfer units, versus 0.038 for the paired estimator and 0.039 for the joint estimator with full-sample censoring. P_{eff} is an empirically calibrated boundary, not a distribution-free constant.
4. **Large-scale audit, external transport, and identification limits** (Sections 6.5, 6.7). Across 10 TCGA cohorts, 3 model families, and 13,500 configurations, we audit the three defects, detect regimes with no effective support, and test cross-cohort transport. We further establish non-identifiability under latent survival-dependent acquisition and use a cross-fitted natural-missingness audit to assess sensitivity under real missingness.

Our stance is grounded in machine-learning evaluation: *without a reliable measurement layer, modality-contribution estimates can be misleading, along with the model comparisons, ablations, and testing decisions that rely on them.*

Figure 1 summarizes the resulting framework, from selective acquisition and censoring to paired, support-aware, adaptive estimation.

2 Related Work

Paired comparison, censoring, and transport. Comparisons between two scores are paired when both are evaluated on the same subjects. DeLong et al. [1] established the canonical comparison for correlated AUCs, and Kang et al. [2] extended it to correlated concordance indices under right censoring. Uno et al. [4] and Liu et al. [22] studied censoring-aware discrimination, while Li et al. [23] studied transport of a fixed predictor. These works motivate pairing, censoring correction, and target-population definition, but do not estimate the paired incremental value of a selectively observed modality.

Multimodal survival and incomplete views. Multimodal cancer models improve fusion and robustness using pathology, genomics, and clinical data [24; 25; 27; 29; 30]. Informative presence can arise from workflow, resources, or disease severity [5]; ICYM^{2}I [3] shows how missingness can distort multimodal informativeness. Other methods impute or reconstruct missing views, or learn shared representations [35; 36; 13; 14; 12; 10]. These methods address model-side robustness, not unbiased population-level valuation of acquiring one modality under selective acquisition and right censoring.

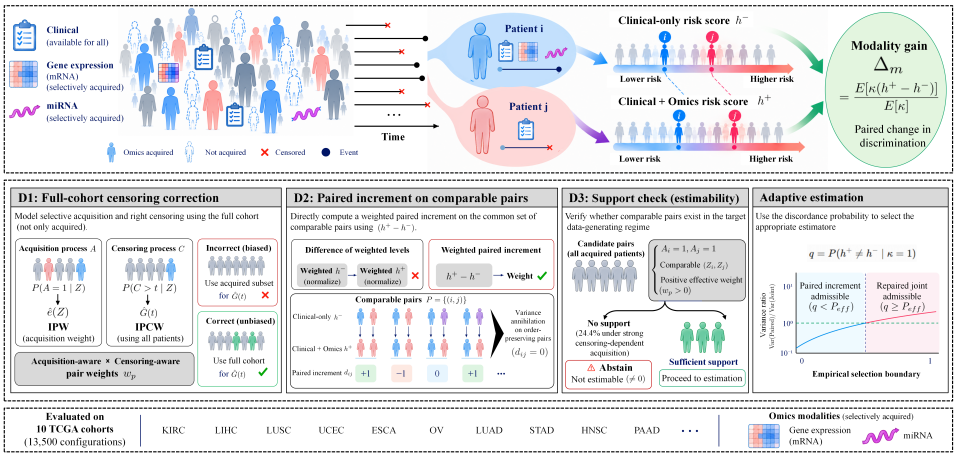

Figure 1: Overview of the evaluation framework. D1 is addressed by full-cohort censoring estimation, D2 by paired increments, and D3 by an explicit support check. The discordance probability q guides empirical estimator selection.

3 Setup, Estimand and Theory

Setup. Let $(X, T, C) \sim P$ with $X = (X_c, X_m)$: the clinical block X_c is always observed; the omics block X_m only when the acquisition indicator $A = 1$. Right censoring gives $(\widetilde{Y}, \delta) = (\min(T, C), \mathbf{1}\{T \leq C\})$. A *frozen* predictor f is fit once on complete training folds. We use the within-predictor controlled-input ablation $\eta^+ = f(x_c, x_m)$ and $\eta^- = f(x_c, 0)$, where 0 is the training-fold reference value in standardized feature coordinates. The estimand therefore quantifies fixed-predictor contribution under this controlled ablation, not intrinsic or causal acquisition value. Write $\kappa_{ij} = \mathbf{1}\{T_i < T_j\}$ and $h^\eta_{ij} = \mathbf{1}\{\eta_i > \eta_j\} + \frac{1}{2}\mathbf{1}\{\eta_i = \eta_j\}$. We abbreviate $h^+ = h^{\eta^+}$ and $h^- = h^{\eta^-}$. The implementation uses strict event-time comparability, so its IPCW kernel is $k^*_{ij} = \delta_i \mathbf{1}\{\widetilde{Y}_i < \widetilde{Y}_j\}/G(T_i)^2$.

Estimand. We define the tie-aware population concordance functional as $\tau_P(\eta) := E[h^\eta_{ij} \mid T_i < T_j] = E[\kappa_{ij} h^\eta_{ij}]/E[\kappa_{ij}]$, and modality value as $\Delta_m := \tau_P(\eta^+) - \tau_P(\eta^-)$. Thus every later identity uses the same score-tie convention. Crucially, τ_P contains no censoring distribution—G enters only estimators—which separates "estimator failure" from "estimand non-identifiability".

Bias decomposition. For acquisition weight $u \geq 0$ with pair weight $u_{ij} = u_i u_j$, define $\Phi(u, \kappa; \eta) = E[u_{ij}\kappa_{ij} h^\eta_{ij}]/E[u_{ij}\kappa_{ij}]$. An exact identity holds, $\Phi - \tau_P = E[(u_{ij} - 1)(h^\eta_{ij} - \tau_P)\kappa_{ij}]/E[u_{ij}\kappa_{ij}]$, and deforming the two observation processes along an acquisition path and a censoring path yields an exact inclusion–exclusion decomposition $B_{\text{sel}} + B_{\text{cens}} + B_{\text{int}}$ in path-integral form (Appendix A.2). Two consequences drive the paper: (i) $B_{\text{sel}} = 0$ exactly when the relevant pair-weighted covariance vanishes; constant acquisition probability is a sufficient special case—selection on *fully observed* X_c, which suffices for regression under conditional MAR, still reweights the law of X and hence changes a population functional: *valuation is not coefficient estimation*; (ii) $B_{\text{int}} \neq 0$ once acquisition depends on censoring information, and one-sided corrections cannot remove it.

Joint identification. Under the full-state acquisition and independent-censoring conditions stated in Appendix A.3, with overlap $e(Z) \geq e_0 > 0$ and $G(t) \geq g_0 > 0$, the jointly weighted estimator (weight $1/[e(Z)G(T)^2]$) converges to τ_P. Its corollary is that one-sided corrections generally converge to different targets. Censoring-only correction reaches an acquisition-weighted concordance. It equals τ_P only when the relevant pair-weighted covariance between acquisition probability and concordance vanishes; constant acquisition probability is a sufficient special case. Acquisition-only IPW reaches a G^2-tilted concordance. It equals τ_P only when the corresponding censoring-weight covariance vanishes; constant G or suitable independence between concordance and event time is sufficient but not necessary.

Non-identifiability. There exist two laws whose acquisition mechanisms both depend on latent T, whose *observed* laws coincide (an explicit mirror construction), yet whose Δ_m differ. A finite-sample illustration (n=5000 $\times$ 10 reps) gives statistically indistinguishable observed laws (KS $p = 0.31$–1.00) while the corresponding truths are -0.102 and $+0.101$. In this finite-sample illustration, each of the four evaluated estimators returns nearly the same mean value under the two laws, and its observed maximum error across the two truths is at least 0.747. More generally, the separation between the two truths implies a minimax lower bound equal to half that separation; the larger observed error reflects the behavior of the evaluated estimators (formal construction in Appendix A.3; finite-sample illustration in Appendix A.7). Under latent-T-dependent acquisition only interval identification or sensitivity analysis remains.

Paired-increment identity and weight-noise annihilation. While pair-level AIPW is detailed in Appendix A.4, the primary methodological advance lies in directly contrasting paired increments. Since $\tau_P(\eta) = E[\kappa h^\eta]/E[\kappa]$ shares a common denominator, subtraction yields the exact *identity* $\Delta_m = E[\kappa_{ij} d_{ij}]/E[\kappa_{ij}]$ with $d_{ij} = h^+_{ij} - h^-_{ij}$, and the two estimable forms are $\hat{\Delta}^{\mathrm{HT}} = \sum k^* \hat{w}\, d / \sum k^*$ (the Horvitz–Thompson form [19]) and $\hat{\Delta}^{\mathrm{Háj}} = \sum k^* \hat{w}\, d / \sum k^* \hat{w}$ with $\hat{w}_{ij} = A_i A_j/(\hat{e}_i \hat{e}_j)$ truncated at τ_{cap}. The consistency statement is for the oracle or consistently estimated, uncapped nuisance version; finite-sample experiments clip weights for stability, as detailed in Appendix A.5.

Proposition 2 (pair-kernel weight-noise comparison)**.** *Let $w_{ij} = A_i A_j/[e(Z_i)e(Z_j)]$ be the oracle acquisition pair weight. Under the stated conditional independence and finite-second-moment conditions, the pair-level weight-noise component for a kernel g is $k^*_{ij}(w_{ij} - 1)g_{ij}$. If $\mathrm{Var}(w_{ij} \mid Z)$ is approximately constant over the relevant support, its conditional second moment is proportional to $E[k^{*2}_{ij} g^2_{ij}]$. Hence the pair-kernel ratio for $g = d$ versus $g = h^+$ is $R_{\mathrm{kernel}} = E[k^{*2} d^2]/E[k^{*2}(h^+)^2] \approx q/P(h^+ \neq 0)$, up to tie terms and the stated dependence approximation. This is a variance proxy, not the asymptotic variance ratio of the full order-2 U-statistics. The latter also contains shared-subject covariance, self-normalisation, nuisance-estimation, and finite-sample truncation terms.*

On pairs whose ordering the modality does not change, $d = 0$ and the weight noise vanishes *exactly*, however large $\hat{w}$ is. Measured on our 13,500 configurations: $q = 0.221$, $P(h^+ \neq 0) = 0.599$, predicted ratio 0.368, observed 0.442 (the excess reflects that $\hat{w}$ and d are weakly dependent through Z; we state the proposition as an approximation, not an equality).

Corollary 1 (adaptive precision and an estimable boundary)**.** *Under the simplifying pair-kernel model, the weight-noise contribution grows with q. Empirically, total estimation error follows the same monotone trend over the audited range and reverses at high q. The quantity q is estimable without observing the complete-fold oracle by $\hat{q} = \sum k^* \hat{w}\, \mathbf{1}\{h^+ \neq h^-\} / \sum k^* \hat{w}$, and the proposition predicts that the paired form's advantage decreases in q—a falsifiable prediction we test in Section 6.5.*

4 THREE SOURCES OF EVALUATION ERROR

D1: censoring estimated on the acquisition-selected subsample. A natural implementation of a jointly weighted Uno-C discards zero-weight (non-acquired) subjects and *then* estimates the censoring Kaplan–Meier. The observed survival pair $(\widetilde{Y}, \delta)$ is available for every evaluation subject, but acquisition is outcome-related, so estimating $\widehat{G}$ on the acquired subset can change the censoring baseline; the independent-censoring condition is defined for the full evaluation sample. Accordingly, $\widehat{G}$ is estimated on the full evaluation sample and indexed by pairs.

D2: a contrast computed as a difference of levels. The field's gain, and the natural corrections, compute $[\sum k^* \hat{w} h^+ / \sum k^* \hat{w}] - [\sum k^* h^- / \sum k^*]$: two different denominators, two different pair sets. This is still consistent, but it places the randomness of $\hat{w}$ on the *level* h^+ rather than on the *increment* d. We therefore use the paired-increment estimator of Section 3. In the no-mask limit, the paired form reproduces Uno-$C(\eta^+)$ − Uno-$C(\eta^-)$ with a maximum absolute difference of 4.7×10^{-16} across all 13,500 configurations. The paired formulation preserves the target estimand while changing its finite-sample computation.

D3: empirical support can vanish, yet the pipeline still reports a number. h^+ is defined only on acquired subjects, so every tier's comparable-pair set is drawn from the acquired subsample.

Under strong censoring-dependent acquisition that subsample can contain *no* comparable pair at all—no event whose later-time comparator is also acquired—so that fold has no empirical support for estimating Δ_m. The population estimand remains well-defined. This occurs in 24.4% of configurations under that mechanism, compared with $\leq 0.6\%$ under every other one (Section 6.6). The unweighted-denominator forms can nevertheless return a numerical value: the Horvitz–Thompson numerator has an empty effective sample while its denominator remains non-zero, yielding exactly 0, whereas the AIPW returns a small non-zero value (median $|\cdot| = 0.0035$ against a median true $|\Delta_m|$ of 0.030). Such unsupported folds are therefore treated as non-estimable rather than assigned a numerical modality value. We require the support condition $\sum_{ij} k^*_{ij} \hat{w}_{ij} > 0$, abstain when it is not satisfied, and report the abstention rate as a diagnostic alongside the estimate.

Candidate estimators and valuation framework. T1 naive observed-only (the field default); T2 complete-case+IPCW; T3 acquisition IPW (the ICYM2I-style observed-training corrector, with the evaluation-side recipe in T3b); T4 joint IPW×IPCW (direct implementation of the joint-identification result); **T4′** = T4 with full-evaluation-sample censoring estimation (D1); T5 Level-Augmented AIPW (standard doubly robust baseline, Appendix A.4); **P** = paired increment incorporating full-evaluation-sample censoring estimation and paired weighting (D1–D2; HT and Hájek normalisation); plus a *zero rule* (always report $\hat{\Delta} = 0$) as a reference testing whether an estimator recovers true signal. Three diagnostics are reported throughout: effective sample size, weight-truncation rate, and covariate balance.

5 Experimental Setup

Data. 10 TCGA cohorts (unified Xena GDC snapshot): KIRC n=510, LIHC 361, LUSC 465, UCEC 533, ESCA 161, OV 369, LUAD 490, STAD 351, HNSC 494, PAAD 176. Clinical block always observed; modality $m \in \{\text{gex}, \text{miRNA}\}$ with in-fold MAD top-400 selection. Protocol: stratified 5-fold CV by event, in-fold selection/scaling/imputation, 3 mask seeds, 3 model families (Cox, RSF, DeepSurv), and $13{,}500$ configurations.

Mechanisms. Acquisition masks on test folds (training folds complete): MCAR; MAR-clinical (age/stage); risk-dependent (clinical-only Cox score); survival-dependent (observed time); censoring-dependent (censoring indicator). Strengths target acquisition rates $0.75/0.55/0.35$; realised rates and all parameters are logged.

Nuisance estimation and stabilisation. On each evaluation fold, a logistic regression estimates $e(Z)$ from all evaluation subjects. Its four covariates are the standardised age–stage composite, the training-fold clinical Cox score, standardised observed time, and the censoring indicator. The propensity model is not cross-fitted. Estimated propensities are clipped to $[0.05, 1]$ for subject-level estimators. Their stabilised weights are capped at the within-configuration 99th percentile. The paired estimator instead clips propensities to $[0.1, 1]$ and caps $A_i A_j/(\hat{e}_i \hat{e}_j)$ at 10. The censoring Kaplan–Meier is fitted on the full evaluation fold and floored at 0.1. Appendix A.4 gives the AIPW cross-fitting and the reduced propensity variants. All asymptotic claims concern the oracle or consistently estimated uncapped version.

Metrics and reporting. The complete-fold oracle/reference Δ_m is computed before applying the acquisition mask. We report (i) $|\hat{\Delta} - \Delta|$ as *block means* over cohort×model blocks with cluster-bootstrap 95% CIs, every column including the zero rule using the same aggregation; (ii) the through-origin **calibration slope** $\sum \Delta\hat{\Delta} / \sum \Delta^2$, since a valuation number is compared across modalities and cohorts and a systematic scale factor drives over- or under-investment; (iii) sign-flip rate; and (iv) Kendall τ against the complete-fold oracle, the ranking use of a contribution column. Unless stated the stratum is strength=strong, mechanism≠MCAR, restricted to the common support of all tiers (Section 6.6) and split at $|\Delta| = 0.05$ (Table 1).

Robustness index. Threshold selection, natural missingness, and propensity misspecification are assessed with frozen/nested validation, a cross-fitted same-model ablation, and full-fold balance diagnostics, respectively (Appendices A.8, A.10, and A.11).

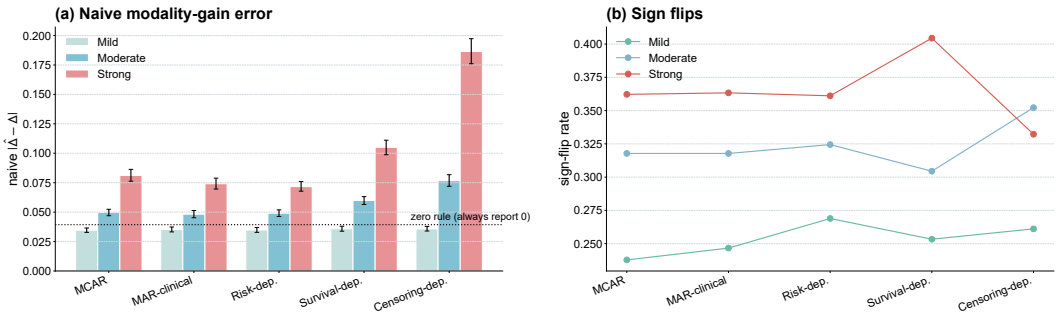

Figure 2: Naive modality-gain estimation under selective acquisition (10 TCGA cohorts, 3 model families; mean with bootstrap 95% CI). (a) Error grows with selectivity and is largest for censoring/survival-coupled mechanisms; the dotted line is the zero rule. (b) Sign-flip rate of the reported value.

Table 1: Estimator ladder. Block means over cohort×model blocks, cluster-bootstrap 95% CIs, strong selective mechanisms excluding MCAR. "Slope" is the through-origin calibration slope (1 = correct scale); τ is Kendall's correlation between estimate and the complete-fold oracle, i.e. how well the tier recovers the *ordering* of modality value. Every column uses the same *common empirical support*, excluding the 226 of 3600 configurations with no supported acquired pair (defect D3, Section 6.6). The first two error columns correspond respectively to the strata n=870 / 27 blocks and n=2504 / 30 blocks.

Estimator	$\lvert\Delta\rvert > 0.05$	$\lvert\Delta\rvert \leq 0.05$	Slope	Sign flip	τ
T1 naive	0.103 [0.092, 0.116]	0.107	0.97	26.0%	0.223
T2 complete-case + IPCW	0.080 [0.068, 0.094]	0.072	0.95	19.8%	0.297
T3 acquisition IPW	0.120 [0.110, 0.131]	0.124	1.03	28.2%	0.198
T4 joint IPW×IPCW	0.120 [0.110, 0.131]	0.121	**1.01**	27.7%	0.206
T4′ + full-sample censoring	0.110 [0.100, 0.120]	0.114	**1.01**	26.3%	0.223
T5 pair-level AIPW	0.076 [0.069, 0.084]	0.063	0.65	23.3%	0.273
P paired increment, HT	**0.054 [0.049, 0.061]**	0.033	0.45	**19.4%**	**0.300**
P paired increment, Hájek	0.080 [0.069, 0.091]	0.076	0.97	**19.4%**	0.289
zero rule (always 0)	0.081	**0.021**	0.00	—	0.000

6 Results

6.1 The phenomenon

Figure 2 shows the same structure across all three model families: error grows with selectivity, and censoring- and survival-coupled mechanisms are worst. In the signal-present stratum, the zero rule has lower block-mean error than the naive estimate (0.081 vs. 0.103 $[0.092, 0.116]$). The naive estimate flips the sign of the value in 26.0% of that stratum.

6.2 Estimator comparison and component analysis

Table 1 gives four readings (the full plots are in Appendix A.6).

(i) Before incorporating D1 and D2, the theoretically motivated corrections perform poorly on both error and rank recovery. T3 and T4 (0.120 each) are *worse* than naive (0.103), which is why a published corrector transplanted here degrades results (Section 6.4). Rank recovery is an independent criterion and agrees: τ falls from 0.223 for naive to 0.198 and 0.206, so weighting damages even the ranking of modalities against one another. Because T4 implements the joint-identification result, its degradation is attributable to finite-sample behavior under the stated identification assumptions.

(ii) D1 and D2 provide distinct empirical improvements. D1 alone moves T4 to 0.110, significant in all four experiment families (Appendix A.6). Adding D2 reaches 0.080, and the paired form

beats naive by $+0.024$ (20/27 blocks, p=2.7×10^{-4}, CI $[+0.013, +0.035]$), T4 by $+0.040$ (25/27, p=4.9×10^{-7}) and T4′ by $+0.030$ (22/27, p=5.5×10^{-5}).

(iii) The paired contrast principle clarifies the in-cohort behavior of complete-case. T2 complete-case, T5 Level-Augmented AIPW, and the paired Hájek form all achieve closely matched block-mean errors around 0.076–0.080 (mutually indistinguishable on in-cohort error, $p > 0.4$). Indeed, complete-case matches the paired form across all four in-cohort axes (0.080 vs 0.080, slope 0.95 vs 0.97, sign flips 19.8% vs 19.4%, τ 0.297 vs 0.289). This parity offers an important conceptual insight: complete-case analysis preserves valid patient-level pairing on the acquired subsample, naturally circumventing the variance inflation of unpaired differencing. Both methods decisively outperform naive evaluation ($+0.024$, $p = 2.7 \times 10^{-4}$). However, as established in the wrong-target corollary in Appendix A.3: complete-case targets the acquired subpopulation $\tau_{P(\cdot|A=1)}$ rather than the population estimand τ_P; its fundamental vulnerability emerges under distribution shift and external transport (Section 6.5).

The candidate estimators separate distinctly in scale calibration and external transfer: Level-Augmented AIPW exhibits slope 0.65 and the un-normalised HT form 0.45, understating true modality value by 35% and 55% due to regression shrinkage when augmenting levels instead of increments. In contrast, the Hájek paired form is near-perfectly calibrated (0.97) while achieving the minimum sign-flip rate (19.4%). While the HT form attains the lowest absolute error (0.054) and highest ranking correlation ($\tau = 0.300$), its uncalibrated scale makes it suitable for ranking but unsuited for quoting absolute value.

(iv) Where the complete-fold oracle is negligible, the zero rule is unbeatable (0.021), the variance floor of the regime—but it attains $\tau = 0$ by construction and cannot rank modalities at all, so error parity with it is not indifference.

6.3 Isolating defect D1

Changing only the censoring-KM estimation set from the acquired subset to the full evaluation fold reduces block-mean error by $+0.008$–$+0.014$ across four experiment families. The gain is largest when acquisition is sparse, as expected when a subsample KM is most biased; the full ablation and acquisition-rate stratification are reported in Appendix A.6. In external transfer, where acquisition is fixed at 0.55 and target cohorts are larger, the effect is small and not significant (p=0.25).

6.4 Misspecification stress, and a strict comparison with ICYM2I

Misspecification. Under an intentionally uninformative propensity model, the paired Hájek estimate changes little, whereas level-augmented estimators are more sensitive. On order-preserving pairs the increment is zero whatever the weights are, but ESS alone can be misleading; full-fold covariate balance provides an additional warning (Appendix A.11).

ICYM2I, observed-training evaluation-side transplant. We also implement the strict observed-training transplant supported by the released survival interface: the model is trained on acquired training samples and evaluated with IPW. It does not recover the target and can increase error; the exact protocol and numerical comparison are given in Appendix A.7. This is a structural limitation of transplanting an evaluation-side corrector to censored survival, not a claim about a fully weighted Cox loss.

6.5 Transport, and the proposition's boundary

We evaluate 10 cross-cohort pairs $\times$ 2 selectivity strengths (20 units): the two official SurvBoard pairs (TCGA-LIHC→ICGC LIRI-JP, TCGA-PAAD→ICGC PACA-CA) and their reverses, a cross-platform pair (TCGA-BRCA→METABRIC) and its reverse, and same-organ pairs (COAD↔READ, LUAD→LUSC, KIRC→KIRP). The frozen predictor is trained once on the source and transferred without retraining; ICGC PACA-AU is excluded because its Illumina probe identifiers cannot be harmonised to symbol/ENSG space without an external annotation table.

Three findings (per-unit detail in the transport figure in Appendix Figure 5).

(i) The paired estimator's advantage decreases with $\hat{q}$ across the 20 transfer units. Regressing the paired form's advantage on $\hat{q}$ over the 20 units, with bootstrap clustered by transfer pair, gives slope

Table 2: Transport: target-cohort $|\hat{\Delta} - \Delta|$ over 20 units. The frozen rule uses $P_{\text{eff}} = 0.3056$ learned on the semi-synthetic pillar and is applied to transfer without using transfer errors.

Estimator	Mean	Median	P90	Worst	> 0.08
P paired increment, Hájek	0.038	0.037	0.058	0.085	2/20
T4′ + full-sample censoring	0.039	0.034	0.070	0.079	0/20
Frozen q-adaptive rule, $P_{\text{eff}} = 0.306$	**0.034**	**0.031**	0.070	**0.079**	**0/20**
oracle {paired, T4′}	0.030	—	—	—	—

-0.085 with 95% CI $[-0.162, -0.054]$, Spearman -0.74, and a zero-advantage crossing at $\hat{q} \approx 0.36$. At $\hat{q} = 0.02$ the paired form attains 0.006 against 0.032 for T4′; at $\hat{q} = 0.50$ it is 0.083 against 0.041. And $\hat{q}$ is estimable: $\text{corr}(\hat{q}, q) = 0.996$ with mean absolute deviation 0.014 at the cohort level (at fold level the correlation drops to $\approx 0.5{-}0.6$, so $\hat{q}$ should be used at aggregate scale).

(ii) Complete-case exhibits severe tail vulnerability under distribution shift. Its mean is 0.045, but it attenuates the target value by 53% on LIHC→LIRI-JP and by 136% on the reverse direction—where the reported value changes sign—and gives the worst single unit of any tier (0.120). Naive exceeds 0.08 in 6 of 20 units; neither the joint estimator with full-sample censoring nor the frozen q-adaptive rule does. Mean error alone understates the tail vulnerability of the naive and complete-case estimators.

(iii) The frozen empirical rule yields a mean absolute error of 0.0340 *on transfer.* The threshold learned from the semi-synthetic pillar is $P_{\text{eff}} = 0.3056$. This improves on the fixed paired choice (0.0378) and the fixed joint estimator with full-sample censoring (0.0390). Within-transfer selection and nested leave-one-pair-out thresholding are reported as exploratory secondary analyses in Appendix A.8. These analyses are exploratory and do not establish a distribution-free threshold. The directional advantage is consistent across pairs, while the finite number of transfer clusters limits power for tail metrics.

6.6 Empirical selection boundary on two pillars, with support checked separately

Section 6.5 tested the *sign* of the proposition's prediction. We now test its *functional form*, on the quantity that decides the comparison. Decomposing each unit's advantage into bias^2 and sd^2 parts, the q-dependence sits entirely in the variance: its slope is significant on both pillars (transfer -0.0064, CI $[-0.0196, -0.0028]$; semi-synthetic -0.159, CI $[-0.187, -0.108]$), while the bias part carries no significant q-dependence on the transfer pillar and is 3.5% of the advantage on the semi-synthetic one.

The simplified proposition concerns the pair-kernel weight-noise contribution, which scales with $q/P(h^+ \neq 0)$ and predicts an idealized slope of $1/P(h^+ \neq 0) = 1.67$ under strict independence. In realistic multi-omics settings where weights and risk scores share latent clinical dependencies through Z, the functional form incorporates a positive covariance inflation factor (Corollary 2.4). Fitting this linear relationship through the origin yields remarkably consistent empirical slopes (Figure 3a): 3.27 $[2.67, 3.88]$ across 239 semi-synthetic blocks (10 cohort clusters) and 3.23 $[2.43, 3.85]$ across the 20 transfer units (10 pair clusters). Both intervals exclude the idealized independence benchmark $1/P(h^+ \neq 0) = 1.67$, confirming the need to account for feature co-dependence. Crucially, inverting this slope defines an empirical selection boundary: $P_{\text{eff}} = 0.306\,[0.258, 0.374]$ on semi-synthetic cohorts and $0.310\,[0.260, 0.411]$ on external transfer. This replication yields an empirical candidate criterion: select the paired form when $\hat{q} < P_{\text{eff}} \approx 0.31$, while treating the threshold as transportable evidence rather than a distribution-free constant (Appendix A.8). It is close to the leave-one-pair-out empirical threshold $\tau_{\text{rule}} = 0.40$ and bracketed by the four empirical crossings (0.285, 0.301, 0.358, 0.399). The variance reversal is directly observed across empirical $\hat{q}$ bins (Figure 3b): the ratio rises monotonically from 0.11 to 2.29, with the optimal estimator transitioning in the bin containing P_{eff}.

Defect D3 completes the procedure. Under strong censoring-dependent acquisition the acquired subsample contains no comparable pair in 24.4% of configurations ($220/900$), against 0.6% for MAR-clinical, 0.2% for MCAR and 0.1% for survival-dependent acquisition (Appendix A.7), across 9 of 10 cohorts and both modalities—a property of the regime, not of a cohort. Those folds have no

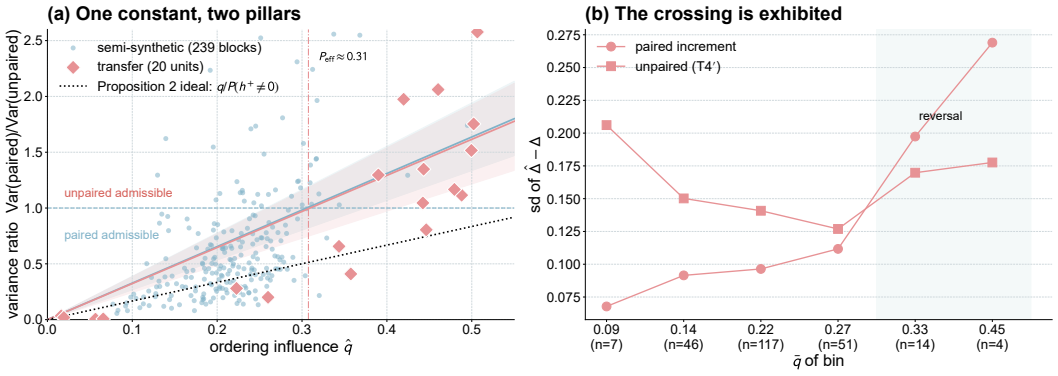

Figure 3: (a) The variance ratio against $\hat{q}$ on two independent pillars (dots: semi-synthetic blocks; diamonds: transfer units; bands: cluster-bootstrap CIs of the through-origin slope, which nearly coincide); the dotted line is the proposition's idealised $q/P(h^+ \neq 0)$, which both bands exclude. (b) The crossing exhibited.

empirical support for estimating Δ_m, yet the Horvitz–Thompson form returns exactly 0 in all 226 unsupported configurations in the main strong non-MCAR audit and the AIPW a small non-zero number. Gating on $\sum k^* \hat{w} > 0$ turns this into a stated abstention, making the valuation a three-way decision—paired increment, unpaired joint, or abstain—and the abstention rate a reportable diagnostic of the acquisition design.

6.7 Natural missingness

SurvBoard's natural missingness provides a real-world sensitivity audit, although the mechanism is unknown. A cross-fitted audit uses the same frozen predictor with and without the miRNA block on LIHC, KIRC, and LUSC. Estimator behavior is heterogeneous across queues: DR is best on KIRC/LIHC and IPW on LUSC (Appendix A.10). These results show that modality valuation is sensitive to the missingness correction under real missingness, without establishing a universal ranking across estimators.

7 Limitations

(1) The complete-fold oracle is observable only in semi-synthetic benchmarks; natural audits rely on estimated propensities, while external transfer uses parametric acquisition because raw ICGC missingness is too sparse for non-parametric fitting. (2) The empirical threshold $P_{\text{eff}} \approx 0.31$ reflects finite-sample covariance in high-dimensional features; Corollaries 2.4–2.5 explain its inflation over the idealized constant, while high-q curvature warrants broader validation. (3) With 10 transfer clusters, tail improvements are empirical patterns rather than large-sample significance; monotonic variance scaling in q and cross-pillar replication of P_{eff} provide structural backing. (4) Discordance $\hat{q}$ is accurate at the cohort level ($\text{corr} = 0.996$) but more variable at fold level, so selection should be study-scale. Finally, semi-parametric efficiency bounds for censored paired functionals remain open, as level-based augmentations risk finite-sample shrinkage.

8 Conclusion

Current multimodal survival benchmarks assess modality contribution through an unpaired difference of separate discrimination metrics, evaluated against a distorted censoring baseline and prone to silent support failures. These three structural defects can distort multimodal benchmark conclusions. By reframing modality valuation as a paired incremental contrast, weight noise is eliminated on concordant pairs, and q motivates an empirically calibrated, non-universal selection boundary between paired and unpaired estimation. Fixing the measurement layer improves the fidelity of multimodal benchmark conclusions and cross-cohort modality valuation.

REPRODUCIBILITY STATEMENT

Complete assumptions, proofs, and estimator derivations are in Appendices A.0–A.5.

Appendices A.6–A.7 isolate D1 and D3. Appendices A.8, A.10, and A.11 cover frozen-threshold validation, cross-fitted natural-missingness analysis, and propensity-balance diagnostics. Logged traces, fixed seeds, and recomputation scripts are in the release package; all tables and figures are recomputable without retraining.

REFERENCES

[1] DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 44(3):837–845, 1988.

[2] Kang, L., Chen, W., Petrick, N. A. & Gallas, B. D. Comparing Two Correlated C Indices with Right-Censored Survival Outcome: A One-Shot Nonparametric Approach. *Statistics in Medicine* 34(4):685–703, 2015.

[3] Choi, Y. S., Jeanselme, V., Elias, P. & Joshi, S. ICYM^{2}I: The Illusion of Multimodal Informativeness under Missingness. *International Conference on Learning Representations*, 2026.

[4] Uno, H., et al. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Statistics in Medicine* 30:1105–1117, 2011.

[5] Sisk, R., et al. Informative Presence and Observation in Routine Health Data: A Review of Methodology for Clinical Risk Prediction. *JAMIA* 28(6):1279–1288, 2021.

[6] Di, D., et al. Generating Hypergraph-Based High-Order Representations of Whole-Slide Histopathological Images for Survival Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 1–16.

[7] Fan, L., Sowmya, A., Meijering, E. & Song, Y. Cancer Survival Prediction From Whole Slide Images With Self-Supervised Learning and Slide Consistency. *IEEE Transactions on Medical Imaging* 42(5):1401–1412, 2023.

[8] Han, X., Zhou, H., Tian, Z., Du, S. & Gao, Y. Inter-Intra Hypergraph Computation for Survival Prediction on Whole Slide Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47(7):6006–6021, 2025.

[9] Liu, C., et al. Asymmetric Graph-Guided Multitask Survival Analysis With Self-Paced Learning. *IEEE Transactions on Neural Networks and Learning Systems* 33(2):654–666, 2022.

[10] Liu, J., et al. One Model to Synthesize Them All: Multi-Contrast Multi-Scale Transformer for Missing Data Imputation. *IEEE Transactions on Medical Imaging* 42(9):2577–2591, 2023.

[11] Luo, H., et al. An Efficient Survival Prediction Framework Based on Multimodal Multi-Instance Fuzzy Optimal Transport. *IEEE Transactions on Fuzzy Systems* 33(11):4055–4068, 2025.

[12] Meng, X., et al. Multi-Modal Modality-Masked Diffusion Network for Brain MRI Synthesis With Random Modality Missing. *IEEE Transactions on Medical Imaging* 43(7):2587–2598, 2024.

[13] Ning, Z., et al. Relation-Aware Shared Representation Learning for Cancer Prognosis Analysis With Auxiliary Clinical Variables and Incomplete Multi-Modality Data. *IEEE Transactions on Medical Imaging* 41(1):186–198, 2022.

[14] Reza, M. K., Prater-Bennette, A. & Asif, M. S. Robust Multimodal Learning With Missing Modalities via Parameter-Efficient Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47(2):742–754, 2025.

[15] Shtossel, O., Koren, O. & Louzoun, Y. Left Barrier Loss for Unbiased Survival Analysis Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47(12):11270–11282, 2025.

[16] Tang, L., et al. Intervention-Based Mixup Augmentation for Multiple Instance Learning in Whole-Slide Image Survival Analysis. *IEEE Transactions on Medical Imaging* 45(1):98–111, 2026.

[17] Wang, D., et al. Deep Multimodal Complementarity Learning. *IEEE Transactions on Neural Networks and Learning Systems* 34(12):10213–10224, 2023.

[18] Yang, H., Sun, J. & Xu, Z. Learning Unified Hyper-Network for Multi-Modal MR Image Synthesis and Tumor Segmentation With Missing Modalities. *IEEE Transactions on Medical Imaging* 42(12):3678–3689, 2023.

[19] Horvitz, D. G. & Thompson, D. J. A Generalization of Sampling Without Replacement from a Finite Universe. *JASA* 47:663–685, 1952.

[20] Chernozhukov, V., et al. Double/Debiased Machine Learning for Treatment and Structural Parameters. *Econometrics Journal* 21(1):C1–C68, 2018.

[21] Wissel, D., Janakarajan, N., Grover, A., Toniato, E., Rodríguez Martínez, M. & Boeva, V. SurvBoard: Standardized Benchmarking for Multi-omics Cancer Survival Models. *Briefings in Bioinformatics* 26(5):bbaf521, 2025.

[22] Liu, W., Liu, D. & Zhang, Z. Evaluating Prognostic Biomarkers for Survival Outcomes Subject to Informative Censoring. *Statistical Methods in Medical Research* 33(8), 2024.

[23] Li, H., et al. Estimating the Area under the ROC Curve When Transporting a Prediction Model to a Target Population. *Biometrics* 79(3):2382–2395, 2023.

[24] Chen, R. J., Lu, M. Y., Wang, J., Williamson, D. F. K., Rodig, S. J., Lindeman, N. I. & Mahmood, F. Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. *IEEE Transactions on Medical Imaging* 41(4):757–770, 2022.

[25] Li, Z., Jiang, Y., Lu, M., Li, R. & Xia, Y. Survival Prediction via Hierarchical Multimodal Co-Attention Transformer: A Computational Histology-Radiology Solution. *IEEE Transactions on Medical Imaging* 42(9):2678–2689, 2023.

[26] Shao, W., Liu, J., Zuo, Y., Qi, S., Hong, H., Sheng, J., Zhu, Q. & Zhang, D. FAM3L: Feature-Aware Multi-Modal Metric Learning for Integrative Survival Analysis of Human Cancers. *IEEE Transactions on Medical Imaging* 42(9):2552–2565, 2023.

[27] Hou, W., Lin, C., Yu, L., Qin, J., Yu, R. & Wang, L. Hybrid Graph Convolutional Network With Online Masked Autoencoder for Robust Multimodal Cancer Survival Prediction. *IEEE Transactions on Medical Imaging* 42(8):2462–2473, 2023.

[28] Zhou, H., Zhou, F. & Chen, H. Cohort-Individual Cooperative Learning for Multimodal Cancer Survival Analysis. *IEEE Transactions on Medical Imaging* 44(2):656–667, 2025.

[29] Wang, Z., Zhang, Y., Xu, Y., Imoto, S., Chen, H. & Song, J. Histo-Genomic Knowledge Association for Cancer Prognosis From Histopathology Whole Slide Images. *IEEE Transactions on Medical Imaging* 44(5):2170–2181, 2025.

[30] Jiang, S., Cai, L., Gan, Z., Wang, Y., Tang, G. & Zhang, Y. Uncertainty-Aware Survival Analysis With Dirichlet Distribution for Multi-Scale Pathology and Genomics. *IEEE Transactions on Medical Imaging* 45(2):554–568, 2026.

[31] Yang, S., Zhou, H., Yang, Y., Wang, W., Chen, S., Yang, G., Fu, H. & Zhu, L. LCM-Net: LLM-Driven Cross-Modality MoE Feature Fusion Network for Cancer Survival Analysis. *IEEE Transactions on Medical Imaging* 45(5):2505–2517, 2026.

[32] Yan, R., Zhang, X., Jiang, Z., Wang, B., Bian, X., Ren, F. & Zhou, S. K. Pathway-Aware Multimodal Transformer: Integrating Pathological Image and Gene Expression for Interpretable Cancer Survival Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 48(1):896–913, 2026.

[33] Qiu, Y., Jiang, K., Yao, H., Wang, Z. & Satoh, S. Does Adding a Modality Really Make Positive Impacts in Incomplete Multi-Modal Brain Tumor Segmentation? *IEEE Transactions on Medical Imaging* 44(5):2194–2205, 2025.

[34] Yang, Y., Pan, H., Jiang, Q.-Y., Xu, Y. & Tang, J. Learning to Rebalance Multi-Modal Optimization by Adaptively Masking Subnetworks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47(6):4553–4566, 2025.

[35] Zhang, Y., Peng, C., Wang, Q., Song, D., Li, K. & Zhou, S. K. Unified Multi-Modal Image Synthesis for Missing Modality Imputation. *IEEE Transactions on Medical Imaging* 44(1):4–18, 2025.

[36] Zeng, P., Yang, M., Lu, Y., Zhang, C., Hu, P. & Peng, X. Semantic Invariant Multi-View Clustering With Fully Incomplete Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46(4):2139–2150, 2024.

AI-USE STATEMENT

LLM-based assistants (including coding agents) were used for experiment pipeline implementation and debugging, numerical verification of derivations, and drafting and translation assistance. All mathematical statements were derived and verified by the authors; all numbers were regenerated from logged per-sample data, and every table and figure is recomputable from the released CSVs. No human participant data were newly collected.

A PROOFS

A.0 Preliminaries and assumptions. Subjects are i.i.d.; pairwise estimators are order-2 U-statistics over pairs (i, j). Pairs sharing a subject are dependent, so the expectation identities below are pair-kernel identities, and consistency follows from the corresponding U-statistic laws of large numbers. Define $h_{ij}^{\eta} = \mathbf{1}\{\eta_i > \eta_j\} + \frac{1}{2}\mathbf{1}\{\eta_i = \eta_j\}$, and abbreviate $h^{+} = h^{\eta^{+}}$, $h^{-} = h^{\eta^{-}}$, and $d_{ij} = h_{ij}^{+} - h_{ij}^{-}$. The observation process maps (X, T, C, A) to $(Z, A, A{\cdot}X_m, \widetilde{Y}, \delta)$.

For identification, Z denotes variables observed for every evaluation subject and available to the acquisition-propensity model. It need not be independent of the censoring process. Let $e(Z) = P(A = 1 \mid Z)$. We use the following assumptions:

(A1) Acquisition. The acquisition mechanism satisfies $E[A/e(Z) \mid X, T, C, Z] = 1$, with $e(Z) \geq e_0 > 0$ on the target support. A sufficient formulation is $A \perp (X, T, C) \mid Z$. Thus acquisition may depend on observed censoring or survival information through Z, but not on residual latent information after conditioning on Z.

(A2) Censoring. $C \perp (X, T)$ and $G(t) = P(C \geq t) \geq g_0 > 0$ on the relevant event-time support.

(A3) Nuisance and implementation. The nuisance estimators for e and G are consistent, pair support has positive probability, and the oracle or consistently estimated *uncapped* weights are used for the asymptotic claim. Finite-sample clipping is an implementation safeguard and need not be exactly unbiased.

A.1 LEMMA 0 (IPCW IDENTITY) AND LEMMA 1 (EXACT WEIGHTED IDENTITY)

Lemma 0. Under $C \perp (X, T)$ and $G(t) = P(C \geq t) > 0$, with $k_{ij}^{*} = \delta_i \mathbf{1}\{\widetilde{Y}_i < \widetilde{Y}_j\}/G(T_i)^2$ we have $E[k_{ij}^{*} g_{ij}] = E[\kappa_{ij} g_{ij}]$ for any g measurable w.r.t. (X, T).

Proof. If $T_i < T_j$, then on $\delta_i = 1$ we have $\widetilde{Y}_i = T_i$, and $\widetilde{Y}_j > T_i$ exactly when subject j remains under observation beyond T_i. Hence $E[\delta_i \mathbf{1}\{C_j \geq T_i\} g_{ij}/G(T_i)^2] = E[\mathbf{1}\{T_i < T_j\} g_{ij} G(T_i) G(T_i)/G(T_i)^2] = E[\kappa_{ij} g_{ij}]$, using $E[\delta_i \mid T_i] = G(T_i)$ and independence of both censoring times from (X, T). If $T_i \geq T_j$, the strict comparable-pair indicator is zero on both sides. Score ties are handled separately by the half-credit definition of h^{η}.

Consequently $\hat{\tau}^{\mathrm{IPCW}}(\eta) = \Phi(1, k^{*}; \eta) \to \tau_P(\eta)$. Weighting both ends of the pair would additionally introduce a $G(T_j)$ factor; Lemma 0 therefore places $1/G(T_i)^2$ on the *event* subject only.

Lemma 1. $\Phi(u, \kappa; \eta) - \tau_P(\eta) = E[(u_{ij} - 1)(h_{ij}^{\eta} - \tau_P(\eta))\kappa_{ij}]/E[u_{ij}\kappa_{ij}]$.

Proof. Subtract τ_P times the denominator from the numerator and use $E[\kappa h] = \tau_P E[\kappa]$.

Corollary (MCAR unbiasedness). If $A \perp (X, T)$ so that $e \equiv e_0$, then u_{ij} is constant and independent of (h, κ), and the numerator vanishes.

A.2 EXACT THREE-TERM DECOMPOSITION

Along an acquisition path $e_\varepsilon = 1 + \varepsilon(e - 1)$ and a censoring path κ_γ ($\gamma = 0$: full comparability; $\gamma = 1$: realised censoring), let $b(\varepsilon, \gamma) = \Phi(e_\varepsilon, \kappa_\gamma) - \Phi(1, \kappa)$. Then the *exact* identity $b(1, 1) - b(0, 0) = B_{\mathrm{sel}} + B_{\mathrm{cens}} + B_{\mathrm{int}}$ holds with $B_{\mathrm{sel}} = b(1, 0) - b(0, 0)$, $B_{\mathrm{cens}} = b(0, 1) - b(0, 0)$ and $B_{\mathrm{int}} = b(1, 1) - b(1, 0) - b(0, 1) + b(0, 0)$, and each term admits a path-integral form by the fundamental theorem of calculus: $B_{\mathrm{sel}} = \int_0^1 \partial_\varepsilon b(s, 0)\, ds$, $B_{\mathrm{cens}} = \int_0^1 \partial_\gamma b(0, t)\, dt$, $B_{\mathrm{int}} = \int_0^1 \int_0^1 \partial_{\varepsilon\gamma}^2 b(s, t)\, ds\, dt$. Point derivatives alone do *not* sum to the finite change for this ratio functional, because $e_{\varepsilon,i} e_{\varepsilon,j}$ is quadratic in ε; a numerical check shows point-derivative sums can miss about 30% of the total. Consequences: (i) $B_{\mathrm{sel}} = 0$ iff the symmetric covariance of $e(X)$ with h^{η} given κ vanishes, which constant e implies; (ii) $B_{\mathrm{cens}} = 0$ under independent censoring with a consistent k^{*} (Lemma 0); (iii) if A depends on δ, B_{int} contains the integrated cross moment, nonzero whenever e has nonzero slope in δ and h is correlated with δ, and unless $B_{\mathrm{sel}} = B_{\mathrm{cens}} = 0$ one-sided corrections cannot cancel it.

A.3 THEOREM 1 (JOINT IDENTIFICATION), ITS COROLLARY, AND THEOREM 2

Theorem 1. Under (A1)–(A3), with $u_i = A_i/e(Z_i)$ and $k = k^*$, the numerator satisfies $E[(A_iA_j/e_ie_j)k^*_{ij}h^+_{ij}] = E[\kappa_{ij}h^+_{ij}]$. Condition on the full state $(X_i, T_i, C_i, Z_i, X_j, T_j, C_j, Z_j)$. By (A1), $E[A_i/e_i \mid X_i, T_i, C_i, Z_i] = 1$ and likewise for j. Hence the acquisition factors integrate out even though k^*_{ij} contains the censoring realization. The remaining expectation is $E[k^*_{ij}h^+_{ij}]$, to which Lemma 0 applies under (A2). Therefore $E[(A_iA_j/e_ie_j)k^*_{ij}h^+_{ij}] = E[k^*_{ij}h^+_{ij}] = E[\kappa_{ij}h^+_{ij}]$. The denominator follows by the same argument with $h \equiv 1$, and the ratio converges by (A3). The η^- term needs no acquisition weight. *Transport remark:* mechanism stability is not part of identification; transporting source-side nuisances to a target requires it, otherwise target-side nuisances must be re-estimated, which is the design of the main-text transport section.

Corollary (wrong targets). (i) With acquisition selection left uncorrected, the resulting functional is weighted by e_ie_j. Equality with τ_P holds exactly when the corresponding pair-weighted covariance between e_ie_j and h, conditional on comparability, vanishes. Constant acquisition probability is a sufficient special case. (ii) With acquisition IPW but without the G-correction, an observed comparable pair survives with probability $P(\delta_i{=}1, \widetilde{Y}_j{>}T_i \mid T_i{<}T_j) = G(T_i)^2$ (event subject uncensored, comparison subject still at risk at T_i). The resulting target is the G^2-tilted concordance

$$\frac{E[h\,\mathbf{1}\{T_i < T_j\}G(T_i)^2]}{E[\mathbf{1}\{T_i < T_j\}G(T_i)^2]}.$$

It equals τ_P exactly when the corresponding covariance with $G(T_i)^2$ vanishes. Constant G, or suitable conditional independence between h and T_i on comparable pairs, is sufficient but not necessary. (iii) Acquisition may depend on observed censoring information through Z under (A1). If that information is omitted from Z, the acquisition moment condition fails and the residual includes the interaction term B_{int}.

Theorem 2 (explicit mirror construction). Take Z constant, $f(X_c, x_m) = x_m$ (so $\eta^+ = X_m$, $\eta^- = 0$), $C \equiv 150.5$ (so $\delta = \mathbf{1}\{T \le 150\}$), support $T \in \{1, \dots, 300\}$ split into event region $E = \{1, \dots, 150\}$ and censored region $R = \{151, \dots, 300\}$ with $P(T \in E) = P(T \in R) = \frac{1}{2}$ in both laws, and $s(t) = (t - 225.5)/149 \in [-1, 1]$, noting $s(451 - t) = -s(t)$. On E both laws coincide: $T|E \sim \text{Unif}\{1, \dots, 150\}$, $X_m|E \sim N(0, 1)$, $A = \mathbf{1}\{U \le \frac{1}{2}\}$ with $U \sim \text{Unif}[0, 1]$ independent of all else. On R: $T|R \sim \text{Unif}\{151, \dots, 300\}$; $P^{(1)}$: $X_m = +a\,s(T) + \varepsilon$, $A = \mathbf{1}\{U \le \frac{1}{2} - \beta s(T)\}$; $P^{(2)}$: $X_m = -a\,s(T) + \varepsilon$, $A = \mathbf{1}\{U \le \frac{1}{2} + \beta s(T)\}$; with $\varepsilon \sim N(0, 1)$, $0 < \beta \le \frac{1}{2}$, $a > 0$.

(i) Both mechanisms depend on latent T: $P(A = 1|T = t) = \frac{1}{2} - \beta s(t)$ on R is non-constant. *(ii) Observed laws coincide:* the mirror map $\iota : (T, \varepsilon, U) \mapsto (451 - T, \varepsilon, U)$ is a bijection of R onto itself with $s \circ \iota = -s$; pushing $P^{(1)}$'s law of (s, X_m, A) through ι gives $(-s, -as + \varepsilon, \mathbf{1}\{U \le \frac{1}{2} + \beta s\})$, exactly $P^{(2)}$'s law. E is law-identical by construction, and $\widetilde{Y}, \delta$ are functions of the region and, on E, of T. Hence the laws of $(A, A{\cdot}X_m, \widetilde{Y}, \delta)$ coincide. *(iii) Δ_m differs:* (E, E) pairs are law-identical; (E, R) pairs have equal concordance in both laws because $S = s(T_j)$ is symmetric ($aS \overset{d}{=} -aS$); (R, R) comparable pairs carry the difference: for $s_j > s_i$, $P^{(1)}$'s $X^j_m - X^i_m = +a(s_j - s_i) + (\varepsilon_j - \varepsilon_i)$ is positive with probability $> 1/2$ while $P^{(2)}$'s is $< 1/2$; as $a \to \infty$ these concordances tend to 0 and 1 respectively, and their conditional mass is at least $1/3$. Since $\eta^- = 0$ makes $\tau_P(\eta^-) = 1/2$ in both laws, $\Delta_m(P^{(1)}) \ne \Delta_m(P^{(2)})$.

Corollary (interval identification). The identification set $\{\tau_{P'} : P' \text{ observationally consistent with } P\}$ has width increasing in the censored mass and in the sensitivity of the acquisition mechanism to t, which is the parameterisation basis for sensitivity analysis.

A.4 THE PAIR-LEVEL AIPW AND WHY IT ATTENUATES

With $w_{ij} = k^*_{ij}$ and a pair outcome model $\hat{m}^+_{ij}$ targeting $m^+(Z_i, Z_j) = E[h^+_{ij} \mid \kappa_{ij} = 1, Z_i, Z_j]$, define $\Psi^+ = E[w_{ij}\hat{m}^+_{ij} + w_{ij}\frac{A_iA_j}{\hat{e}_i\hat{e}_j}(h^+_{ij}A_iA_j - \hat{m}^+_{ij})]$ and $\hat{\tau}^{\text{DR}} = \Psi^+/\hat{E}[\kappa]$. (i) If $\hat{e} = e$ then $\Psi^+ = E[\kappa h^+]$ for any $\hat{m}$; (ii) if $\hat{m}^+ = m^+$ the second term has conditional expectation zero for any $\hat{e}$, since $E[w(h^+ - m^+) \mid Z] = E[\kappa h^+ \mid Z] - m^+(Z)E[\kappa \mid Z] = 0$. The outcome model $\hat{m}^+$ is

fitted by two-fold cross-fitting [20] so that the augmentation is independent of the fold it corrects. The implementation clips the pair ratio at τ_{cap} for variance control, which trades exactness of the propensity-correct branch for stability.

Empirically this estimator is *attenuated*: its through-origin calibration slope is 0.65 in the signal-present stratum (main-text Table 1), i.e. it understates modality value by 35%. The reason is visible from the identity: the augmentation is fitted to the concordance *level* h^+, whose conditional mean is near $1/2$, so shrinkage of $\hat{m}^+$ towards $1/2$ shrinks $\hat{\tau}^{\text{DR}}(\eta^+)$ towards $1/2$ and hence Δ_m towards 0. The paired-increment construction of A.5 avoids this because the quantity being modelled and weighted is d, which is centred at the estimand itself rather than at $1/2$.

Nuisance estimation and weight stabilisation. For each evaluation fold, $e(Z)$ is estimated with scikit-learn logistic regression (L2 penalty, default $C = 1$, maximum 1000 iterations) fitted on the entire evaluation fold. The full propensity covariate set is $Z = (z_{\text{clinical}}, z_{\text{risk}}, z_{\text{time}}, 1 - \delta)$, where z_{clinical} is the average of standardised age and ordinal stage, z_{risk} is a clinical-only Cox score fitted on the training fold, and z_{time} is standardised observed time. Thus the propensity model is not cross-fitted. For the reduced-covariate diagnostic we use $(z_{\text{clinical}}, 1 - \delta)$; the extreme misspecification variant uses three independent Gaussian covariates. Propensity estimates are clipped to $[0.05, 1]$ for subject-level estimators. Stabilised subject weights are $A_i\bar{A}/\hat{e}_i$ and are capped at the within-configuration 99th percentile. The paired estimator clips propensities to $[0.1, 1]$ and caps $A_iA_j/(\hat{e}_i\hat{e}_j)$ at $\tau_{\text{cap}} = 10$. The censoring survival function is estimated by Kaplan–Meier on the full evaluation fold, with a floor $G \geq 0.1$, rather than on the acquired subset. The AIPW outcome model is fitted with two-fold cross-fitting; e itself is not cross-fitted. Its pair ratio is capped at 10 and its fitting weights at 10^2. Asymptotic claims refer to the oracle or consistently estimated uncapped version.

Acquisition mechanism	Mask-generating variable	Propensity covariates
MCAR	none	full Z
MAR-clinical	z_{clinical}	full Z
Risk-dependent	z_{risk}	full Z
Survival-dependent	z_{time}	full Z
Censoring-dependent	$1 - \delta$	full Z

A.5 Paired-increment identity and variance annihilation

Lemma 2 (paired-increment identity). $\Delta_m = \tau_P(\eta^+) - \tau_P(\eta^-) = E[\kappa_{ij}d_{ij}]/E[\kappa_{ij}]$.

Proof. By definition $\tau_P(\eta) = E[\kappa h^\eta]/E[\kappa]$, and the denominator $E[\kappa]$ does not depend on η. Hence $\Delta_m = (E[\kappa h^+] - E[\kappa h^-])/E[\kappa] = E[\kappa(h^+ - h^-)]/E[\kappa]$.

Corollary 2.1 (estimable forms). Under (A1)–(A3) and Lemma 0, the oracle or consistently estimated, uncapped versions of

$$\hat{\Delta}^{\text{HT}} = \frac{\sum_{ij} k^*_{ij}\hat{w}_{ij}d_{ij}}{\sum_{ij} k^*_{ij}}, \qquad \hat{\Delta}^{\text{Háj}} = \frac{\sum_{ij} k^*_{ij}\hat{w}_{ij}d_{ij}}{\sum_{ij} k^*_{ij}\hat{w}_{ij}}, \qquad \hat{w}_{ij} = \frac{A_iA_j}{\hat{e}_i\hat{e}_j},$$

are consistent for Δ_m. The proof conditions on the full latent state and uses $E[A_i/e_i \mid X_i, T_i, C_i, Z_i] = 1$ from (A1), followed by Lemma 0; the self-normalising denominator is likewise consistent for $E[\kappa]$. Under finite-sample truncation at τ_{cap}, the HT form is biased towards zero while the Hájek form remains approximately calibrated, which we observe: slopes 0.45 and 0.97, respectively.

Equivalence under complete acquisition. In the no-mask limit ($A \equiv 1$, $e \equiv 1$) $\hat{\Delta}^{\text{Háj}}$ reproduces Uno-$C(\eta^+)$ − Uno-$C(\eta^-)$ to a maximum absolute difference of 4.7×10^{-16} across all 13,500 configurations. This numerical equivalence is consistent with the algebraic identity in Lemma 2 and confirms that the paired formulation preserves the target estimand.

Unpaired formulation. The field's computation and the natural corrections evaluate

$$\hat{\Delta}^{\text{unpaired}} = \frac{\sum k^*\hat{w}\,h^+}{\sum k^*\hat{w}} - \frac{\sum k^*\,h^-}{\sum k^*},$$

with two different denominators and two different pair sets. It remains consistent but places the randomness of $\hat{w}$ on the level h^+ instead of the increment d.

Proposition 2 (pair-kernel weight-noise comparison). Let $w_{ij} = A_i A_j / [e(Z_i)e(Z_j)]$ denote the oracle acquisition pair weight. Under the stated conditional independence and finite-second-moment conditions, the pair-level weight-noise component for a kernel g is $k^*_{ij}(w_{ij} - 1)g_{ij}$. If $\mathrm{Var}(w_{ij} \mid Z)$ is approximately constant over the relevant support, its conditional second moment is proportional to $E[k^{*2}_{ij} g^2_{ij}]$. Hence the pair-kernel ratio for $g = d$ versus $g = h^+$ is

$$R_{\mathrm{kernel}} = \frac{E[k^{*2} d^2]}{E[k^{*2}(h^+)^2]} \approx \frac{q}{P(h^+ \neq 0)}, \qquad q := P(h^+ \neq h^- \mid \kappa = 1).$$

Proof. For one pair, decompose $w = 1 + (w-1)$. The weight-noise component is $k^*(w-1)g$, whose conditional second moment is $k^{*2}\sigma^2_w g^2$. The displayed ratio follows after suppressing the common factor. This is a pair-kernel comparison, not an exact total U-statistic variance identity: shared-subject covariance, self-normalisation, nuisance estimation, and truncation are handled empirically below. Finally, $d^2 \in \{0, \frac{1}{4}, 1\}$ and $(h^+)^2 \in \{0, \frac{1}{4}, 1\}$; neglecting tie terms gives the stated mass ratio.

Mechanism. On every comparable pair for which $h^+ = h^-$, we have $d = 0$, so acquisition-weight noise on that pair is annihilated exactly. Therefore, when q is small relative to the support of h^+, the pair-kernel variance proxy favors the paired form. In our experiments $q < P(h^+ \neq 0)$ on average, whereas the empirical advantage reverses in high-q regimes.

Empirical pair-kernel comparison. Over 13,500 configurations, $q = 0.221$ and $P(h^+ \neq 0) = 0.599$, so the proposition predicts 0.368; the measured ratio $\mathrm{Var}[k^*(\hat{w} - 1)d]/\mathrm{Var}[k^*(\hat{w} - 1)h^+]$ is 0.442. The difference is consistent with weak dependence between $\hat{w}$ and d through Z, because the acquisition mechanism is outcome-related while d depends on η^+. Accordingly, Proposition 2 provides an order-of-magnitude approximation rather than an exact equality.

Corollary 2.2 (adaptive precision). Under the simplifying pair-kernel model, the weight-noise contribution grows with q. Empirically, total estimation error follows the same monotone trend over the audited range and reverses at high q. The precision of a low-influence modality is therefore expected to be higher: the precision of the estimator is governed by the quantity it estimates. Measured by quartiles of q on the common support, the paired estimator's standard deviation rises monotonically from 0.087 at $\bar{q} = 0.142$ to 0.140 at $\bar{q} = 0.301$, whereas the unpaired form does not rise at all but drifts down over 0.158–0.136; the two therefore cross in the top quartile, matching the variance reversal in main-text Figure 3(b). In the lowest-q quartile the signal-to-noise ratio $E|\Delta_m|/\mathrm{sd}$ nearly doubles when the increment is weighted rather than the level ($0.18 \to 0.32$), and the advantage shrinks monotonically to parity by the top quartile (0.40 vs 0.40).

Note that q is not a proxy for $|\Delta_m|$: a modality can churn the ordering (large q) while contributing no net gain ($\Delta_m \approx 0$). In our transfer set the pair TCGA-PAAD→ICGC PACA-CA has $q = 0.43$ with $\Delta_m = 0.003$. The pair (q, Δ_m) therefore characterises a modality on two axes, net gain and churn.

Corollary 2.3 (estimating q, and the resulting boundary). q depends on η^+ and is computable only on acquired pairs, so it must be inverse-probability restored:

$$\hat{q} = \frac{\sum_{ij} k^*_{ij} \hat{w}_{ij} \mathbf{1}\{h^+_{ij} \neq h^-_{ij}\}}{\sum_{ij} k^*_{ij} \hat{w}_{ij}}.$$

Across the 10 transfer pairs, $\mathrm{corr}(\hat{q}, q) = 0.996$ with mean absolute deviation 0.014 and maximum 0.052; at fold level the correlation falls to 0.5–0.6, so $\hat{q}$ should be used at cohort or aggregate scale. Proposition 2 yields a falsifiable monotonicity prediction: the advantage of the paired form over the unpaired one should decrease in q. Regressing $\mathrm{adv} = |\mathrm{err}|(\mathrm{T4}') - |\mathrm{err}|(\mathrm{paired})$ on $\hat{q}$ over the 20 transfer units, with bootstrap clustered by transfer pair,

$$\mathrm{adv} = +0.0304 - 0.0850\,\hat{q}, \quad \text{slope 95\% CI } [-0.1624, -0.0542], \quad \text{Spearman} = -0.740,$$

with a zero-advantage crossing at $\hat{q} \approx 0.358$. The slope confidence interval excludes zero, and the estimated advantage crosses zero near $\hat{q} \approx 0.358$.

Corollary 2.4 (dependence inflation; the constant is not 1). Relax Proposition 2 by allowing the conditional weight variance to depend on the covariates, $\mathrm{Var}(\hat{w} \mid Z) = \sigma^2_w(Z)$, and keep everything else. The leading pair-kernel weight-noise contribution becomes proportional to $E[k^{*2}\sigma^2_w(Z)g^2]$, so the pair-kernel ratio is *exactly*

$$R_w = \frac{E[k^{*2}\sigma^2_w d^2]}{E[k^{*2}\sigma^2_w (h^+)^2]} = \frac{\tilde{q}}{\tilde{P}}, \qquad \tilde{q} := \frac{E[k^{*2}\sigma^2_w d^2]}{E[k^{*2}\sigma^2_w]}, \quad \tilde{P} := \frac{E[k^{*2}\sigma^2_w (h^+)^2]}{E[k^{*2}\sigma^2_w]},$$

i.e. the same two masses, tilted by σ_w^2. Writing

$$c := \frac{\tilde{q}/q}{\tilde{P}/P(h^+ \neq 0)} \quad \Longrightarrow \quad R_w = c \cdot \frac{q}{P(h^+ \neq 0)},$$

Proposition 2 is the special case $c = 1$, and $c > 1$ precisely when the weight variance is relatively heavier on the order-*changing* pairs than on the pairs the omics score orders at all. This is what informative acquisition produces: $\hat{w}$ is driven by $e(Z)$ and d by $\eta^+(Z)$, so the two are co-determined by Z and cannot be independent. The direction of the inequality is therefore predicted, not assumed: on the component Proposition 2 isolates we measure $R_w = 0.442$ against the idealised 0.368, i.e. $\hat{c} = 1.20$.

Corollary 2.5 (empirical selection boundary and the effective ordering budget). What decides which form to use is not R_w but the ratio of *total* error variances,

$$R_{\text{est}} := \frac{\text{Var}(\hat{\Delta}^{\text{Háj}} - \Delta_m)}{\text{Var}(\hat{\Delta}^{\text{unpaired}} - \Delta_m)},$$

which additionally carries the randomness of the self-normalising denominator and the effect of truncation. Both forms are approximately calibrated (through-origin slopes 0.97 and 1.01), so the ordering of mean squared errors is approximately the ordering of variances in this audit. The paired form is preferred when $R_{\text{est}} < 1$. Empirically R_{est} is proportional to q through the origin, with the same constant on two independent evidence pillars:

pillar	units	clusters	$\hat{q}$ range	slope	$P_{\text{eff}} = 1/\text{slope}$
semi-synthetic	239	10 cohorts	$[0.077, 0.495]$	$3.27\ [2.67, 3.88]$	$\mathbf{0.306}\ [0.258, 0.374]$
transfer	20	10 pairs	$[0.016, 0.507]$	$3.23\ [2.43, 3.85]$	$\mathbf{0.310}\ [0.260, 0.411]$

Hence the operational candidate criterion

$$\boxed{\text{select the paired form when } \hat{q} < P_{\text{eff}} \approx 0.31.}$$

Two structural aspects clarify this behavior. First, the idealized baseline constant $1/P(h^+ \neq 0) = 1.67$ lies outside both confidence intervals, reflecting the fact that the empirical variance ratio incorporates the covariance factor $c > 1$ (Corollary 2.4, measured at $\hat{c} = 1.20$ on R_w) alongside denominator self-normalization and truncation terms. Second, P_{eff} represents an empirical boundary for this evaluated regime: the theory supplies the governing discordance parameter q and the linear functional form, while the two pillars provide overlapping empirical estimates. The frozen and nested validation in Appendix A.8 supports transportability of the rule but does not establish an invariant or distribution-free threshold.

Direct observation of the variance reversal across empirical bins. Binning the semi-synthetic blocks on $\hat{q}$ with fixed edges (so the sparse high-q end is not compressed by quantiles), the variance ratio rises monotonically through unity and crosses 1 in the bin containing P_{eff}:

$\hat{q}$ bin	blocks	$\bar{q}$	paired sd	unpaired sd	ratio	winner
$[0.00, 0.10)$	7	0.089	0.068	0.206	0.11	paired
$[0.10, 0.18)$	46	0.144	0.092	0.150	0.37	paired
$[0.18, 0.25)$	117	0.215	0.096	0.141	0.47	paired
$[0.25, 0.31)$	51	0.275	0.112	0.127	0.78	paired
$[0.31, 0.40)$	14	0.326	0.197	0.170	1.35	**unpaired**
$[0.40, 1.00)$	4	0.449	0.269	0.178	2.29	**unpaired**

The absolute-error and variance criteria cross their respective selection boundaries in the same bin. The two top bins contain 14 and 4 blocks, respectively, so the reversal is supported by the monotone trend together with the clustered slope rather than by those bins alone.

The q-dependence is a variance phenomenon. Decomposing the advantage per unit into bias^2 and sd^2 parts and regressing each on $\hat{q}$: the variance part has a significant negative slope on both pillars (semi-synthetic -0.159, CI $[-0.187, -0.108]$, crossing 0.285; transfer -0.0064, CI $[-0.0196, -0.0028]$, crossing 0.399), whereas the bias part carries no significant q-dependence on the transfer pillar (CI $[-0.0061, +0.0027]$) and accounts for 3.5% of the total advantage on the semi-synthetic one. Proposition 2 therefore governs the term that actually decides the comparison, and the four empirical crossings — 0.285, 0.301, 0.358, 0.399 — bracket P_{eff}.

A.6 Ablating the censoring-estimation set (Defect D1)

Table 3: D1 in isolation: identical acquisition weights, with only the censoring KM estimation set changed. Block means are in the signal-present stratum; the gain is acquired-subset error minus full-fold error, so positive values favor full-fold censoring estimation.

Experiment family	KM on acquired	KM on full fold	Gain	Blocks	p
Main, cohorts 1–5	0.126	0.118	+0.008	11/14	2.3×10^{-3}
Main, cohorts 6–10	0.113	0.101	+0.012	12/13	1.2×10^{-3}
Pair-level truncation	0.127	0.114	+0.014	12/13	4.9×10^{-4}
Estimator variant zoo	0.129	0.117	+0.012	8/9	7.8×10^{-3}

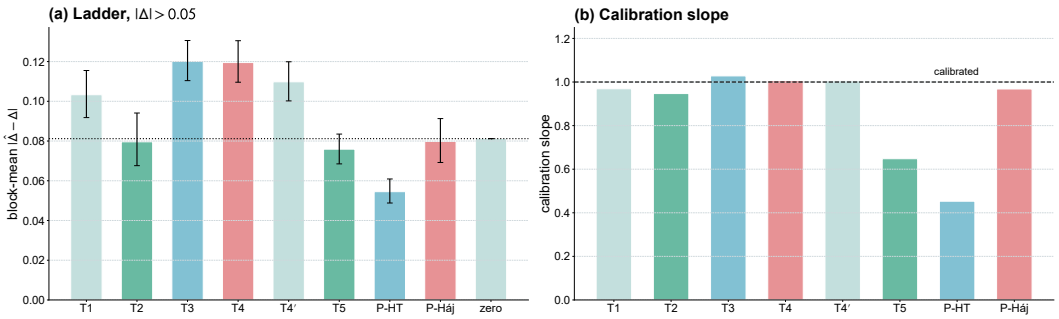

Figure 4: The estimator ladder of main-text Table 1 plotted. (a) Block-mean error by tier in the signal-present stratum with cluster-bootstrap 95% CIs; dotted line is the zero rule. (b) Through-origin calibration slope: the pair-level AIPW (T5) and the un-normalised paired HT form attenuate the reported value, whereas the paired Hájek form does not.

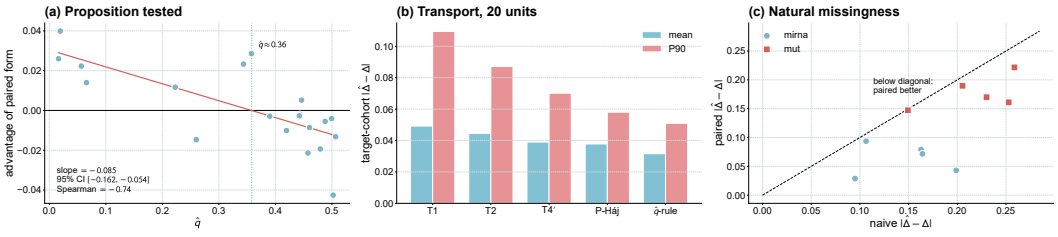

Figure 5: Transport detail. (a) The proposition tested out of sample by sign: the paired form's advantage declines with $\hat{q}$ over 20 transfer units (line: fit; dotted: zero crossing); the stronger functional-form test is main-text Figure 3(a). (b) Mean and P90 target-cohort error; q-rule is the exploratory within-transfer rule (App. A.8). (c) Original natural-missingness audit; estimand-aligned audit in App. A.10.

A.7 Finite-sample support failure and abstention (Defect D3)

The D3 diagnostic is intentionally reported separately from the two-panel empirical-selection figure in the main text. It is a conditional support check rather than a second estimate of the q-boundary: the rate is near zero for most mechanisms but rises to 24.4% under strong censoring-dependent acquisition ($220/900$ configurations). In those configurations the effective pair support is empty, so the paired valuation cannot be estimated on those folds. The diagnostic therefore motivates an explicit abstention rule rather than imputing a numerical zero.

Finite-sample illustration of Theorem 2. We evaluate the mirror construction in Appendix A.3 with n=5000, 10 replicates, a=6, and β=0.45. Both acquisition mechanisms depend on latent T, while their observed laws are statistically indistinguishable (KS $p = 0.31$–1.00 over X_m, acquired-X_m, acquisition rate, and $\widetilde{Y}$). Nevertheless, the corresponding truths are $\Delta(P^{(1)}) = -0.102$ and

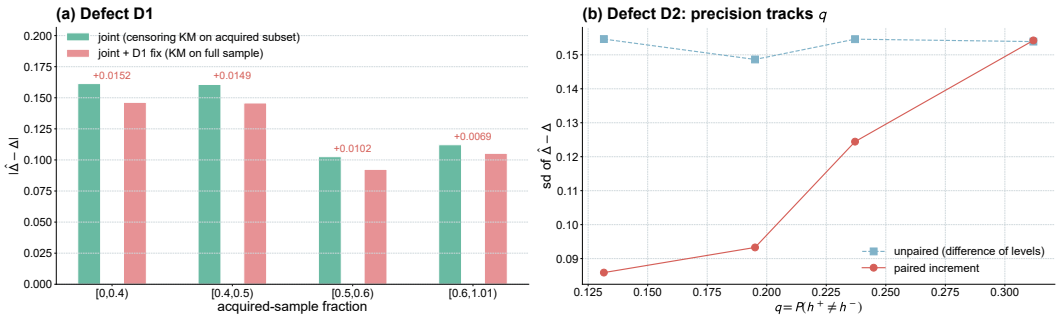

Figure 6: (a) Defect D1: the gain from estimating the censoring distribution on the full sample grows as the acquired fraction falls. (b) Defect D2: the paired estimator's standard deviation tracks the modality's ordering influence q, while the unpaired form is flat—the adaptive-precision corollary 2.2.

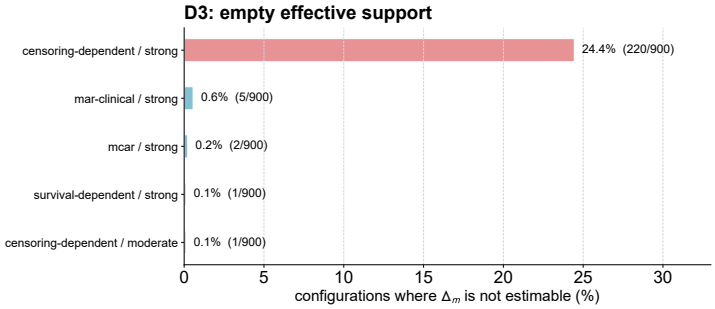

Figure 7: D3 support-failure diagnostic. Bars show the fraction of configurations with empty effective pair support, averaged over the reported acquisition regimes; labels give the number of unsupported configurations over the total. The large rate is concentrated in strong censoring-dependent acquisition, while the other audited regimes are close to zero.

$\Delta(P^{(2)}) = +0.101$. In this finite-sample illustration, each of the four evaluated estimators returns nearly the same mean value under the two laws, and its observed maximum error across the two truths is at least 0.747. More generally, the separation between the two truths implies a minimax lower bound equal to half that separation; the larger observed error reflects the behavior of the evaluated estimators. This construction illustrates the non-identifiability characterized by Theorem 2. Under latent-T-dependent acquisition, the observed data support only interval identification or sensitivity analysis.

A jointly weighted Uno-C implemented by discarding zero-weight subjects and then estimating the censoring Kaplan–Meier evaluates $\widehat{G}$ on the acquired subsample. The observed survival pair $(\widetilde{Y}, \delta)$ is available for every evaluation subject, and (A2) is stated for the full population; when acquisition is outcome-related the acquired subsample can have a different censoring distribution, so $\widehat{G}$ is misscaled. We estimate $\widehat{G}$ once on the full evaluation sample and index it by pair. Isolating this single change (Table 3) improves the joint estimator in all four experiment families, with gains of $+0.008$ to $+0.014$ in block-mean absolute error and $p \leq 8 \times 10^{-3}$. Across acquisition-rate strata, the gain decreases from $+0.0152$ to $+0.0069$ as the acquired fraction increases, showing that the difference is largest when acquisition is scarce (Figure 6). In the transport experiments, where the acquisition rate is fixed at 0.55 and target cohorts are larger, the effect is small and not significant ($p = 0.25$). The empirical benefit of full-sample censoring estimation is therefore concentrated in scarce-acquisition regimes.

ICYM^{2}I comparison protocol. T3b applies the released survival training interface to the observed-training setting. Masks are applied to the training folds, a Cox predictor is fit on the acquired training samples, and evaluation-side IPW is then applied on KIRC/LIHC/ESCA with moderate and strong non-MCAR masks. T3b uses evaluation-side IPW; the Cox training loss itself is not fully weighted. With a complete-trained predictor, naive evaluation gives 0.093 (large-truth stratum 0.104);

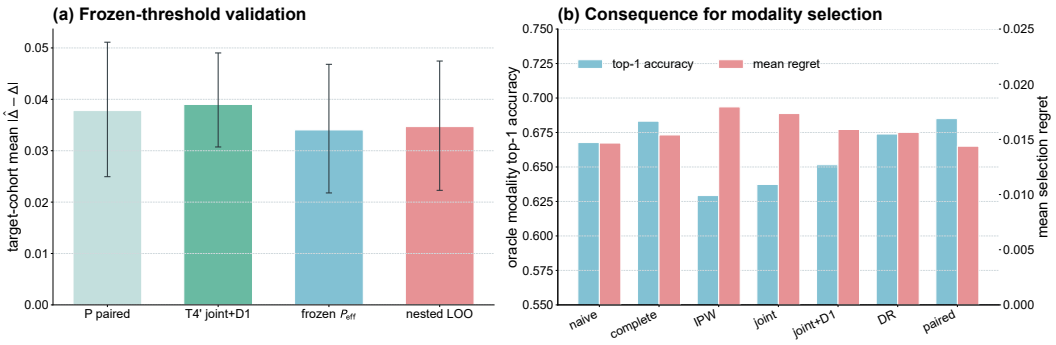

Figure 8: Additional validation experiments. (a) Cross-pillar validation of a threshold learned on semi-synthetic blocks and frozen before external transfer; error bars are pair-cluster bootstrap 95% CIs. (b) Consequences of the estimated contribution column for choosing between gex and miRNA, measured by oracle top-1 selection accuracy and regret on the main strong-selective stratum.

evaluation-side IPW gives 0.108 (0.120); and the observed-training transplant gives 0.120 (0.143). The observed-training transplant has the largest error among these three settings. This pattern is consistent with the wrong-target analysis in the corollary of Appendix A.3 and with the D1/D2 analysis.

A.8 Frozen-threshold validation of the empirical selection rule

The main text reports an adaptive rule that chooses between the paired increment and the joint estimator with full-sample censoring according to the discordance rate $\hat{q}$. To evaluate the rule without using transfer errors for threshold selection, we perform cross-pillar validation in which the threshold is estimated from the semi-synthetic pillar only. First, we estimate the through-origin variance slope on the 239 semi-synthetic blocks only. Its inverse gives $P_{\text{eff}} = 0.3056$, with the semi-synthetic cluster-bootstrap interval $[0.2576, 0.3744]$ reported in the main analysis. We then freeze this value and apply it without tuning to the 20 external-transfer units. The rule selects the paired increment when $\hat{q} < P_{\text{eff}}$ and the joint estimator with full-sample censoring otherwise. For comparison, we also report a fixed threshold of 0.31 and a nested leave-one-pair-out rule. In the nested rule, the threshold is selected from the grid $\{0.10, 0.15, \ldots, 0.50\}$ and its direction is learned using the other nine transfer pairs only.

Table 4: Cross-pillar and nested validation of the $\hat{q}$-adaptive rule on 20 external-transfer units. Intervals are pair-cluster bootstrap 95% CIs for the mean absolute error. No transfer error is used to obtain the frozen P_{eff} value.

Estimator	Mean $\lvert\hat{\Delta} - \Delta\rvert$	95% CI	Median	P90	Worst
P paired, fixed	0.0378	[0.0249, 0.0511]	0.0372	**0.0580**	0.0847
T4′ joint+D1, fixed	0.0390	[0.0307, 0.0490]	0.0338	0.0701	0.0792
Frozen $P_{\text{eff}} = 0.306$	**0.0340**	[0.0218, 0.0468]	**0.0306**	0.0701	**0.0792**
Frozen threshold 0.31	**0.0340**	[0.0218, 0.0468]	**0.0306**	0.0701	**0.0792**
Nested LOO rule	0.0347	[0.0223, 0.0475]	**0.0306**	0.0701	**0.0792**

Table 4 shows that the cross-pillar rule remains useful under cross-pillar threshold selection: the frozen rule reduces mean error from 0.0378 for the paired estimator and 0.0390 for the fixed joint estimator with full-sample censoring to 0.0340. The nested transfer rule gives 0.0347, confirming the direction of the result but also showing that the gain is smaller than the 0.0316 obtained when the threshold is selected from the same transfer units. The selected nested thresholds are mostly 0.40, with one held-out pair selecting 0.25 and one selecting 0.45; this variation is evidence against treating 0.31 as a universal constant. We therefore interpret P_{eff} as an empirically transportable boundary for this data regime, rather than as a distribution-free constant or a universal hyperparameter.

A.9 DOES VALUATION CHANGE MODALITY-SELECTION DECISIONS?

The paper motivates modality valuation as a measurement layer for benchmark comparisons and acquisition decisions. To test this implication directly, we use the existing 10-cohort phenomenon traces without retraining any predictor. For each cohort, model, fold, mechanism, strength and mask seed, we select the modality with the largest estimated gain among gex and miRNA. The oracle choice is the modality with the largest complete-fold Δ_m. We report top-1 selection accuracy and the decision regret

$$R = \max_m \Delta_m - \Delta_{\widehat{m}},$$

where $\widehat{m}$ is the modality selected by the estimator. Regret is zero when the selected modality agrees with the oracle or when the two modality values are tied.

Table 5: Modality-selection consequences on the main strong selective stratum (10 cohorts, non-MCAR mechanisms, 1,613 valid decisions except DR with 1,800 decisions).

Estimator	Valid decisions	Top-1 accuracy	Mean regret
Naive	1,613	0.668	0.0147
Complete-case	1,613	0.683	0.0154
Acquisition IPW	1,613	0.629	0.0179
Joint IPW×IPCW	1,613	0.637	0.0173
Joint+D1	1,613	0.652	0.0159
DR	1,800	0.674	0.0156
Paired Hájek	1,613	**0.685**	**0.0144**

As shown in Table 5, the paired Hájek estimator has the highest top-1 modality-selection accuracy and the lowest mean regret in the main stratum, although the differences are moderate rather than universal. The important result is that valuation error is decision-relevant: even the naive contribution column selects the wrong modality in roughly one third of the decisions, while acquisition-only and joint weighting perform worse than naive on this criterion. This complements the sign-flip and Kendall-τ results: ranking correlation measures a continuous ordering, whereas regret measures the consequence of using the reported column to choose a modality. Because only two modalities are compared and the oracle is available only in the semi-synthetic setting, this experiment supports decision sensitivity, not a claim that the paired estimator dominates for every acquisition policy.

A.10 CROSS-FITTED NATURAL-MISSINGNESS AUDIT WITH A FROZEN MODEL ABLATION

To align the natural-missingness analysis with the estimand used in the main experiments, we conduct a cross-fitted audit on the LIHC, KIRC, and LUSC TCGA queues. In each of five repeated runs, the complete queue is split into five event-stratified folds. The predictor is trained on the training fold only, and the same frozen predictor is evaluated with the miRNA block either retained or zeroed.

Table 6: Cross-fitted natural-missingness audit using one frozen Cox model and the same-model miRNA ablation. Values are mean absolute errors over valid fold-replicates.

Queue	Complete-case	Naive	Acquisition IPW	DR
KIRC	0.0367	0.0346	0.0364	**0.0239**
LIHC	0.0288	0.0552	0.0573	**0.0243**
LUSC	0.0382	0.0265	**0.0214**	0.0295

Table 6 shows a more modest and heterogeneous pattern than the semi-synthetic stress tests. DR has the lowest MAE for KIRC and LIHC, while Acquisition IPW has the lowest MAE for LUSC; no estimator dominates across all three queues. This is the appropriate interpretation of the natural audit: it confirms that the reported valuation can be sensitive to the missingness correction under real missingness, but it does not by itself establish a universal ranking of corrections. Cross-fitting and same-model ablation reduce potential optimism while yielding fewer valid configurations.

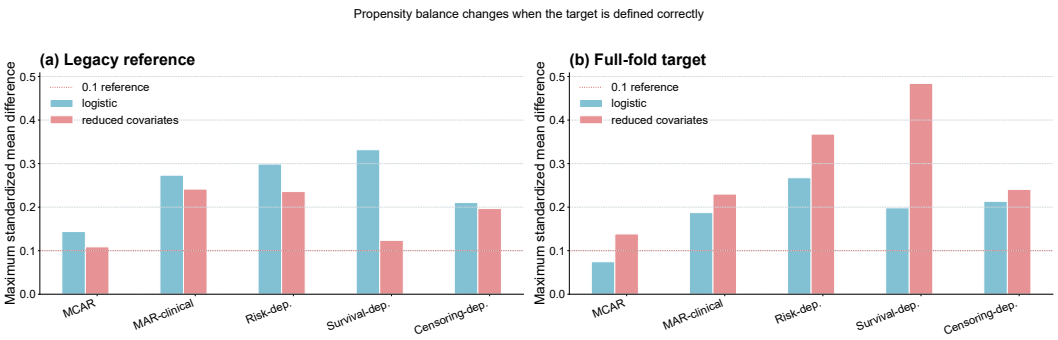

Figure 9: Propensity-balance diagnostics under two reference populations. (a) The acquired sample is used as the reference; (b) the complete test fold is used as the reference. Bars show means over folds, seeds, and strengths for the five semi-synthetic acquisition mechanisms. The difference is most pronounced for the reduced-covariate propensity model and under strong selective mechanisms, supporting the use of balance diagnostics alongside valuation.

A.11 PROPENSITY-BALANCE DIAGNOSTIC UNDER MISSPECIFICATION

The propensity-variant experiment also provides a diagnostic for whether the acquisition model removes the covariate shift that it is intended to remove. We assess propensity balance relative to the complete test-fold population. Specifically, the diagnostic is the maximum standardized mean difference between the inverse-probability-weighted acquired sample and the complete test fold. This reference population directly measures the extent to which weighting recovers the covariate distribution of the target evaluation population.

Table 7: Maximum standardized mean difference, averaged over mild, moderate, and strong selection strengths, folds, seeds, and the five audited cohorts. The target is the complete test-fold covariate mean. Lower values indicate better balance; 0.1 is shown as a conventional descriptive reference.

Acquisition mechanism	Full logistic	Reduced covariates
MCAR	**0.075**	0.138
MAR-clinical	**0.187**	0.230
Risk-dependent	**0.268**	0.368
Survival-dependent	**0.198**	0.484
Censoring-dependent	**0.213**	0.241
All mechanisms	**0.188**	0.292

Table 7 and Figure 9 show that the diagnostic is sensitive to the covariates supplied to the propensity model. The full logistic model has the lowest average imbalance overall (0.188), whereas the reduced model is substantially worse (0.292), especially for survival-dependent selection (0.484). This result is consistent with the intended use of balance as a misspecification warning, but it does not establish that a smaller SMD guarantees a smaller valuation error: the outcome estimator also depends on censoring, overlap and pair support. In particular, under strong risk- and censoring-dependent selection, the SMD remains large even for the full logistic model, indicating substantial residual imbalance despite the availability of an IPW estimator. The diagnostic can therefore serve as an explicit warning for poorly supported acquisition regimes and can inform gating or abstention.