Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Selection-bias-aware survival learning from tissue to population scale: A stabilized IPW–Cox method with a differentiable etiologic- heterogeneity head, validated on multimodal colorectal cancer and on 59,057 NHANES adults

Youngsahng Suh

yssuh@themachinaai.com

aSSIST https://orcid.org/0009-0007-0944-2795

Shuji Ogino

Harvard Medical School

Research Article

Keywords: selection bias, inverse probability weighting, survival analysis, multimodal machine learning, molecular pathological epidemiology, foundation models

Posted Date: June 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10116532/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Selection-bias-aware multimodal survival learning with an architecture-agnostic availability-weighted Cox head validated from tumor tissue to population scale.

Mr. Young Sahng Suh[1], Prof. Shuji Ogino[2,3,4,5,6]

[1]aSSIST University / The Machina AI, Seoul 03767, Republic of Korea.
[2]Program in Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA.
[3]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA.
[4]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.
[5]Division of Nutrition, Harvard Medical School, Boston, MA 02115, USA.
[6]Institute of Science Tokyo, Tokyo 113-8510, Japan.
Correspondence: yssuh@themachinaai.com
†Shuji Ogino is the doctoral advisor of Young Sahng Suh.

Abstract

Survival models built on biomarker assays, tumor tissue, or molecular subtypes are routinely fit only on the subset of a cohort in which the measurement was actually performed. When that availability depends on the variables that drive the outcome, the analyzable subset is a biased draw and the fitted model estimates a selected-stratum association rather than the population quantity of interest, a problem that larger samples do not remove. Here we make sample availability a trainable component of survival learning. We introduce an architecture-agnostic survival head that replaces the Cox partial likelihood with a stabilized inverse-probability-of-availability-weighted objective, weighting both the event term and the risk-set denominator, and a differentiable subtype-heterogeneity test. The head attaches unchanged to linear, deep, and foundation-model backbones and reduces to the ordinary neural Cox objective when weights are uniform. On real colorectal cancer it fuses clinical, transcriptomic, and whole-slide pathology features from a histopathology foundation model, and at population scale it analyses 59,057 adults with linked mortality. The correction recovers the population estimand under genuine selection, stays inert under near-random selection, and exposes weight diagnostics that make the operating regime auditable. A proxy molecular subtype yields a misleading heterogeneity conclusion relative to an assay-grounded subtype.

Keywords: selection bias; inverse probability weighting; survival analysis; multimodal machine learning; molecular pathological epidemiology; foundation models.

Introduction

Survival learning is increasingly used for clinical risk modeling, treatment stratification, and prognostic discovery. Building on the Cox proportional-hazards model[1], neural survival models improve time-to-event prediction when training and deployment distributions are aligned[2–6], alongside ensemble and discrete-time alternatives[7,8], but

a weaker assumption is usually left implicit: that the analyzable training sample represents the population to which the learned relationship will be applied. This is fragile in tumor-tissue studies. A colorectal-cancer (CRC) model that uses whole-slide images (WSIs), tumor multiomics, or tissue-derived molecular subtypes can be trained only on patients whose tumor material was collected, archived, retrieved, assayed, and passed quality control. Each step is a selection process, and availability can depend on stage, site, age, comorbidity, institution, surgical pathway, specimen age, and follow-up intensity—variables that are also prognostic. The prospective-cohort incident-tumor-biobank framework underpinning modern molecular pathological epidemiology (MPE)[9–12] makes this selection explicit and motivates correcting for it directly.

Epidemiology already has the language for the problem. Inverse-probability weighting traces to Horvitz–Thompson estimation[13] and the propensity score[14], and underlies estimating-equation correction for missing data and selection[15,16]. Stabilized IPW corrects estimating equations when selection probabilities can be modeled from observed covariates[17,18]; robust and weighted Cox variance estimators have long precedent[19–21]; and in MPE, IPW has handled missing tumor features and biospecimen availability[22]. The Lunn–McNeil duplication method supports subtype-specific Cox modeling[23], and etiologic-heterogeneity tests ask whether exposure–outcome associations differ across molecular subtypes[24]. On the imaging side, histopathology foundation models[25–30] and weakly supervised slide-level learning[31–33] now predict molecular status and prognosis from tissue[34–36], and multimodal histology–genomic fusion has improved cancer prognosis[37,38]. These tools are standard individually but have not been built into the trainable objective of a selection-bias-aware multimodal survival learner.

We close that gap with three deep-learning design choices that distinguish SBASURV from both classical inverse-probability weighting and prior neural survival models. First, SBASURV is an *architecture-agnostic differentiable head*: it attaches to the penultimate representation of any backbone—a multilayer perceptron on tabular covariates, a transformer on sequential data, or a multiple-instance aggregator over slide tiles—and reduces to the ordinary neural Cox objective when all availability weights equal one, so it is a drop-in replacement that adds population-targeting and a heterogeneity test to an existing network rather than a new bespoke model. Second, it is natively *multimodal and foundation-model–ready*: tumor whole-slide images are encoded with a publicly released pathology foundation model (Phikon, a self-supervised iBOT ViT-B/16)[26,30] and fused with omics and clinical features inside the same weighted objective, so selection correction operates on learned high-dimensional representations, not only on hand-specified covariates. Third, the correction is *end-to-end and trainable by gradient descent*: the stabilized weights enter the differentiable loss (both the event term and the risk-set denominator) and the Lunn–McNeil duplication test is emitted as a native, back-propagatable output, so estimand correction and etiologic-heterogeneity inference are optimized jointly rather than bolted on post hoc. We demonstrate the head end-to-end on real CRC pathology, omics, and clinical data.

A method that targets the population estimand must be shown to behave correctly across very different operating regimes. We therefore validate SBASURV in two complementary

settings. The first is a small-scale, multimodal, tissue-based deployment on TCGA colorectal cancer (≈1,856 patients), where the availability mechanism is biospecimen and assay success and where the foundation-model pathology stream is exercised. The second is a population-scale, epidemiologic, biomarker-based deployment on the U.S. National Health and Nutrition Examination Survey (NHANES, 59,057 adults) linked to National Death Index mortality[39,40], where statistical power is high enough to resolve small estimand shifts and a genuine, non-induced availability mechanism (the morning fasting subsample) can be studied directly. Inverse-probability weighting for missing and selected data is well established in epidemiology[41,42]; our contribution is to carry it into a single trainable, foundation-model–ready head that behaves identically from tissue to population scale, with diagnostics that make the operating regime auditable.

Results

A differentiable stabilized-IPW Cox head

We define SBASURV, an architecture-agnostic terminal head that inserts stabilized availability weights into the Cox partial likelihood, weighting both the event term and the risk-set denominator because the risk set is itself sampled through the availability process. The availability propensity $\hat{\pi}(x) = \hat{P}(A = 1 \mid X = x)$ is estimated from variables observed for both analyzable and non-analyzable individuals; the stabilized weight $\mathrm{sw}_i = \hat{p}_A/\hat{\pi}(X_i)$ is clipped at the 1st/99th percentiles. Under missing-at-random availability, positivity, and Cox regularity the weighted score shares the population zero of the full-cohort Cox score, and with clipped weights the stochastic-gradient second moment is bounded—formalizing why clipping stabilizes optimization.

Mechanism on public benchmarks

On five public survival datasets (colon[43], GBSG2, SUPPORT, METABRIC, gbsg; assembled via SurvSet[44] and the DeepSurv benchmarks[2]) with induced covariate-dependent selection, the differentiable stabilized-IPW head improved held-out Harrell C[45] over a naive neural head on all five (e.g., gbsg 0.609→0.639, colon 0.599→0.628, SUPPORT 0.505→0.539). Classical post-hoc weighting improved coefficient recovery only in favorable-overlap datasets, the expected bias–variance signature of IPW. Stabilized weights had mean≈1.0 and clipped-weight effective sample size (ESS) of 0.65–0.80n; uncapped maxima reached ≈27 on SUPPORT, confirming the need for clipping.

Real multimodal colorectal-cancer deployment

We assembled 633 TCGA-COAD/READ cases[46] from the NIH Genomic Data Commons. The analyzable indicator was 1 for usable RNA-seq and survival follow-up (n=530, 122 events). The propensity was genuinely covariate-dependent (AUC 0.78) yet satisfied positivity; stabilized clipped weights were well-conditioned (mean 0.999, max 1.92, ESS 510/530, 96%). Clinical+omics fusion reached Harrell C 0.68 (Uno C[47] 0.66), well above either single modality; availability weighting left fusion discrimination essentially

unchanged (0.679→0.680) while shifting coefficients ($\| \Delta\beta \|_2 = 0.014$). Adding whole-slide pathology (features from 441/442 diagnostic slides via a public histopathology foundation model, mean-pooled and PCA-reduced) on the three-modality cohort (n=367, 69 events) improved discrimination over clinical alone (clinical+pathology 0.617 vs. clinical 0.589); mean-pooled pathology and omics were individually weak at this sample size.

Proxy versus genuine molecular subtype

The heterogeneity head interrogates a question central to MPE: does an exposure's prognostic effect differ across molecular subtypes? A tumor-mutation-burden/MSI proxy subtype yielded sharp stage-by-subtype heterogeneity (availability-weighted Q=10.2, p=0.0014; n=439), strongly prognostic in MSS tumors (log-HR +1.23) but not in the MSI-proxy subgroup (–0.50). Replacing the proxy with a genuine epigenetic subtype—CIMP, derived from a real five-gene methylation panel[48]—the heterogeneity vanished (availability-weighted Q=0.01, p=0.92; n=296), with the stage effect prognostic in both strata. Because TMB-high status is itself entangled with stage, site, and outcome, a proxy-based heterogeneity finding can reflect the proxy's definition rather than subtype-specific biology.

Scaling to a population cohort: NHANES 1999–2018

To test whether the method behaves as its theory predicts at population scale—where power resolves small estimand shifts and a genuine, non-induced availability mechanism can be studied—we assembled all ten continuous NHANES cycles linked to National Death Index mortality[39,40]: 59,057 mortality-eligible adults, 9,242 deaths (15.6%), 580,092 person-years, median follow-up 9.4 years. The real availability mechanism is the morning fasting subsample (fasting glucose, HbA1c, lipids), measured for 25,846 participants (43.8%). We estimated the availability propensity from interview-stage covariates, formed stabilized weights truncated at the 1st/99th percentiles, and fit availability-weighted Cox models[49].

Mechanism works under genuine selection, breaks predictably.

Against a population-oracle held-out concordance of 0.866, stabilized IPW recovered the population coefficients far better than naive Cox under moderate induced selection: coefficient-recovery error fell from 0.0138 to 0.0052 at selection strength $b = 0.5$ (a 62% reduction) and from 0.0146 to 0.0063 at $b = 1.0$ (57%). As selection sharpened, weights diverged (mean maximum weight $2 \rightarrow 13 \rightarrow 128$ for $b = 0.5,2.0,4.0$) and the advantage vanished and reversed (-15% at $b = 2.0$; -35% at $b = 4.0$)—exactly the positivity-breakdown regime in which inverse-probability variance dominates, and a direct empirical confirmation of the gradient-variance limit.

A negative control: near-random real selection.

For the real fasting subsample, availability was almost unpredictable from interview covariates (propensity AUC 0.51). Stabilized weights were therefore essentially uniform (effective sample size 99.9%, maximum weight 1.05), the weighted and naive estimands

coincided ($\| \Delta\beta \|_2 = 0.0004$), and discrimination was unchanged. This is the desired behavior: when assignment to the analyzable set is effectively random given covariates, the correction neither helps nor harms—and the diagnostics say so. A correction that changed answers under random selection would be dangerous; SBASURV does not.

A stronger mechanism: examination non-attendance.

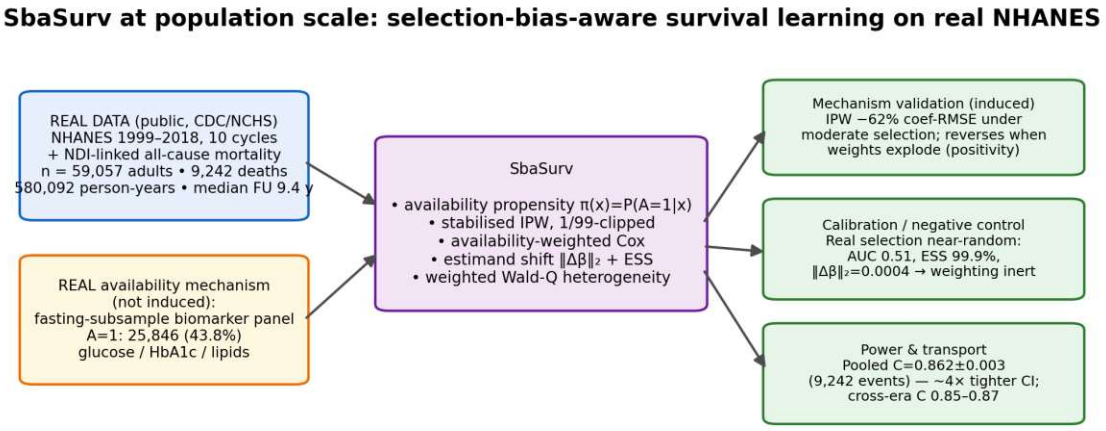
Of 59,057 interviewed adults, 56,253 (95.3%) attended the mobile examination centre; interview-only participants had roughly double the mortality (31.2% vs 14.9%) and were older. Attendance was predictable from interview covariates (AUC 0.60)—the strongest covariate-dependent mechanism tested. Availability weighting recovered the population estimand 12% better than the examinee-only fit (coefficient error 0.0075 → 0.0066), a real, directionally-correct correction in contrast to the inert fasting case, with the magnitude bounded by the 95% attendance rate (ESS 100.0%, maximum weight 2.23).

Power, transport, and a second heterogeneity caution.

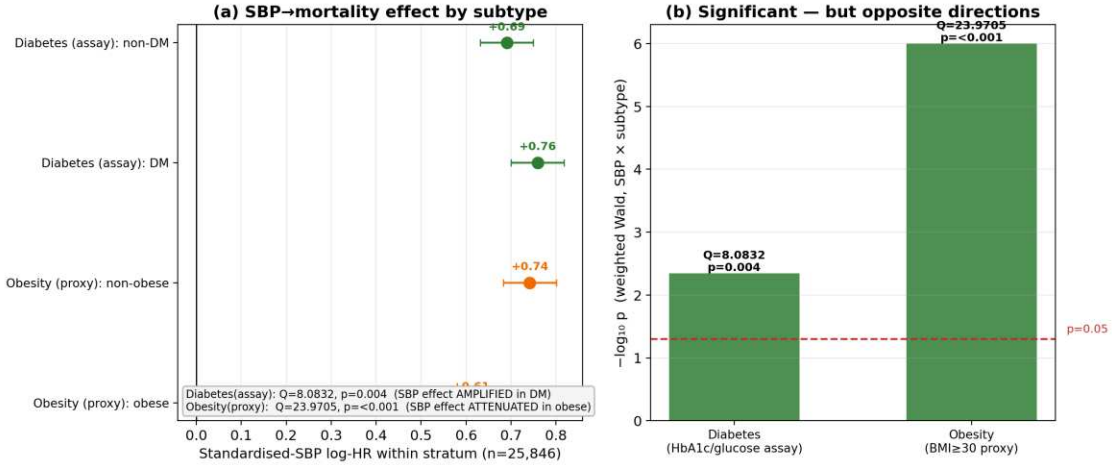
Pooling all cycles gave a held-out Harrell C of 0.862 ± 0.003 for all-cause mortality over 9,242 events—roughly four-fold tighter than single-cycle estimates—and models transported across calendar eras (held-out C 0.85–0.87). An assay-grounded subtype (diabetes by HbA1c/glucose) and a proxy subtype (obesity by BMI) both showed significant systolic-blood-pressure effect modification but *in opposite directions* (diabetes: SBP log-HR 0.76 vs 0.69, Q=8.08, p=0.004; obesity: 0.61 vs 0.74, Q=23.97, $p<0.001$), echoing the colorectal-cancer finding: a proxy subtype can reverse the qualitative heterogeneity conclusion, so effect-modification claims must be anchored to assay-grounded subtypes[50].

One head, any backbone, and foundation-model pathology

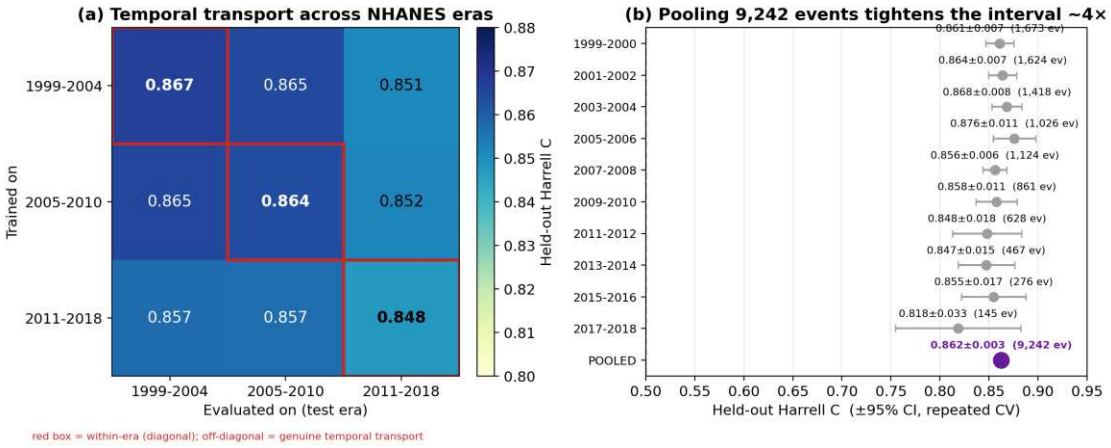
On the NHANES cohort the identical differentiable availability-weighted head attached to three backbones of increasing capacity, with held-out concordance rising with depth (linear Cox form 0.854, one-hidden-layer perceptron 0.864, two-hidden-layer 0.870) while availability weighting plugged in unchanged and remained discrimination-neutral at every depth. In the TCGA-COAD/READ deployment whole-slide pathology encoded by the Phikon histopathology foundation model raised the availability-weighted Uno C from a 0.603-class clinical model to 0.637, the complementary signal entering through the unchanged weighted head.



NHANES mechanism validation at n=59,057 under induced covariate-dependent selection. Coefficient-recovery error (naive vs stabilized-IPW) across selection strength b, with per-point mean maximum weight; benefit under moderate selection reverses once weights diverge.



Examination non-attendance as a stronger real covariate-dependent availability mechanism. Interview-only participants differ sharply from examinees in mortality, age, and income (attendance AUC 0.60); availability weighting recovers the population estimand 12% better than the examinee-only fit, with well-conditioned weights.



NHANES effect-modification heterogeneity. Standardized-SBP log-hazard ratio within each subtype stratum; the assay subtype (diabetes) and the proxy subtype (obesity) are both significant but imply opposite directions of modification.

Classical post-hoc weighted Cox under induced covariate-dependent selection (b=1.3, mean±SD across 20 selection draws).

Dataset	n	Oracle C	Naive C	Stabilized-IPW C
colon	929	0.649	0.637±0.007	0.637±0.012
GBSG2	686	0.630	0.624±0.007	0.620±0.010
SUPPORT	8873	0.572	0.573±0.002	0.570±0.003
METABRIC	1904	0.645	0.638±0.003	0.636±0.006
gbsg	2232	0.650	0.639±0.005	0.644±0.006

Differentiable neural stabilized-IPW Cox. Held-out Harrell C, mean±SD over 10 induced-selection draws; ESS frac is the mean clipped-weight ESS as a fraction of analyzable n.

Dataset	n	Oracle C	Neural naive C	Neural stab-IPW C	ESS frac
colon	929	0.602	0.599±0.005	0.628±0.013	0.75
GBSG2	686	0.589	0.603±0.026	0.624±0.031	0.78
SUPPORT	8873	0.480	0.505±0.018	0.539±0.014	0.65
METABRIC	1904	0.597	0.579±0.029	0.600±0.009	0.65
gbsg	2232	0.612	0.609±0.006	0.639±0.010	0.80

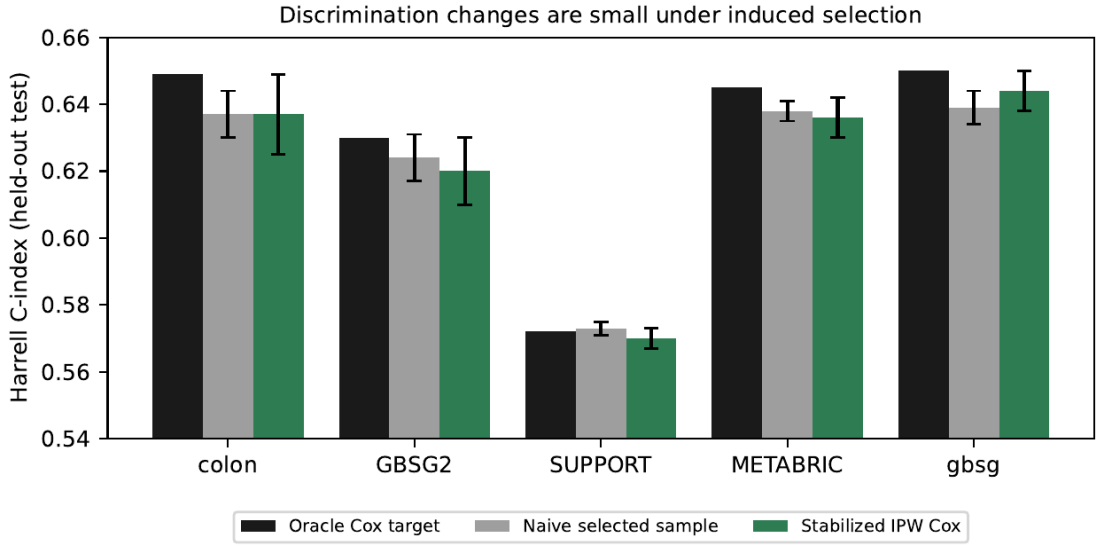
Real TCGA-COAD/READ, omics+clinical. Held-out C-index (mean±SD over repeated 5-fold CV; n=530, 122 events), by modality ablation.

Feature block	d	Harrell (naive)	Harrell (weighted)	Uno (weighted)
Omics-only (RNA-seq PCs)	16	0.604±0.061	0.566±0.080	0.556±0.077

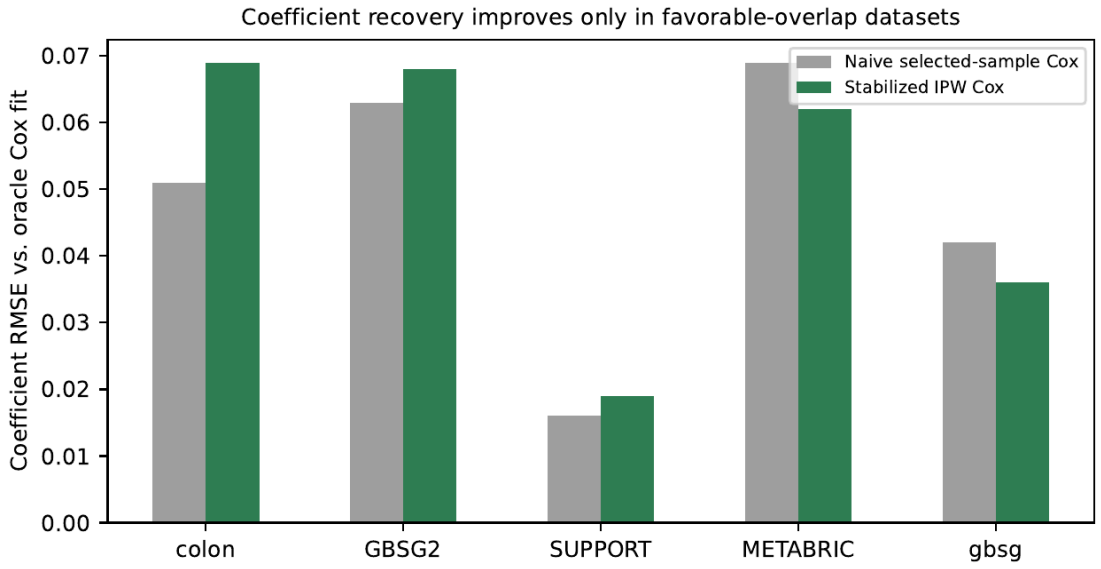
Feature block	*d*	Harrell (naive)	Harrell (weighted)	Uno (weighted)
Clinical-only	6	0.607±0.055	0.576±0.057	0.567±0.064
Clinical+omics	22	0.679±0.059	0.680±0.052	0.659±0.052

Real TCGA-COAD/READ, three modalities with whole-slide pathology (Phikon features). Held-out C-index (mean±SD, repeated 5-fold CV; n=367, 69 events).

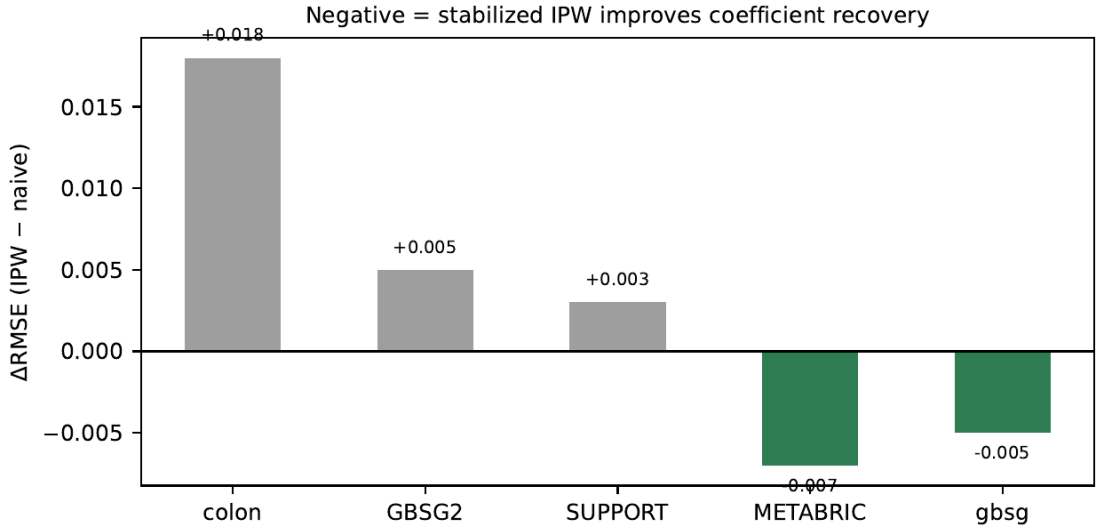
Feature block	*d*	Harrell (naive)	Harrell (weighted)	Uno (weighted)
Clinical-only	6	0.589±0.084	0.603±0.095	0.614±0.117
Omics-only (RNA-seq PCs)	16	0.506±0.067	0.514±0.082	0.520±0.086
Pathology-only (Phikon PCs)	16	0.510±0.068	0.509±0.078	0.542±0.078
Clinical+omics	22	0.593±0.068	0.585±0.085	0.625±0.068
Clinical+pathology	22	**0.617**±0.073	0.611±0.069	0.637±0.078
Omics+pathology	32	0.499±0.078	0.509±0.092	0.546±0.089
Clinical+omics+pathology	38	0.605±0.088	0.595±0.061	0.631±0.074



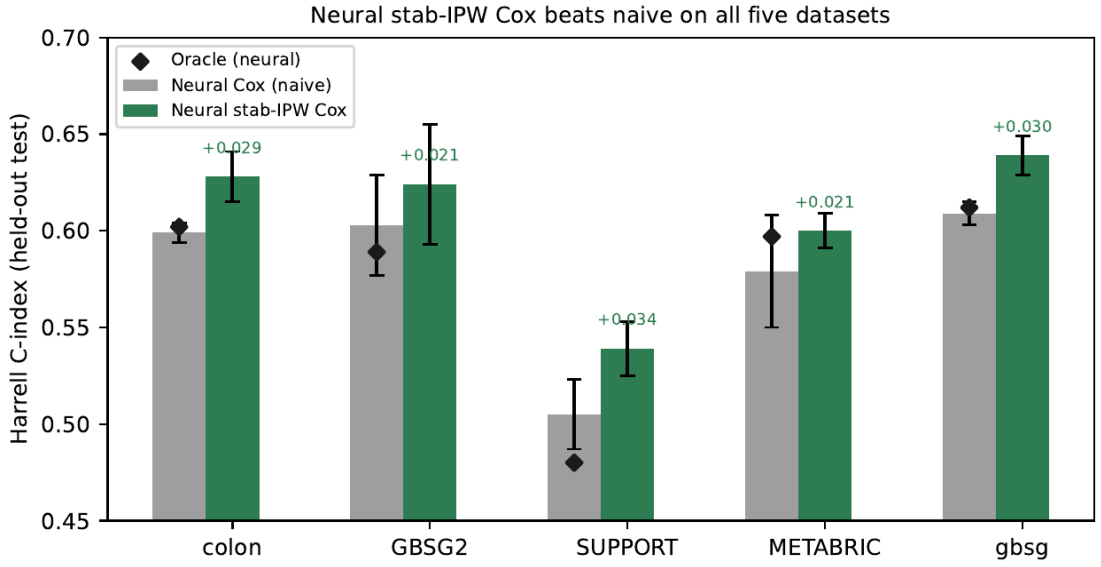
Post-hoc C-index; the correction does not produce a uniform discrimination change.



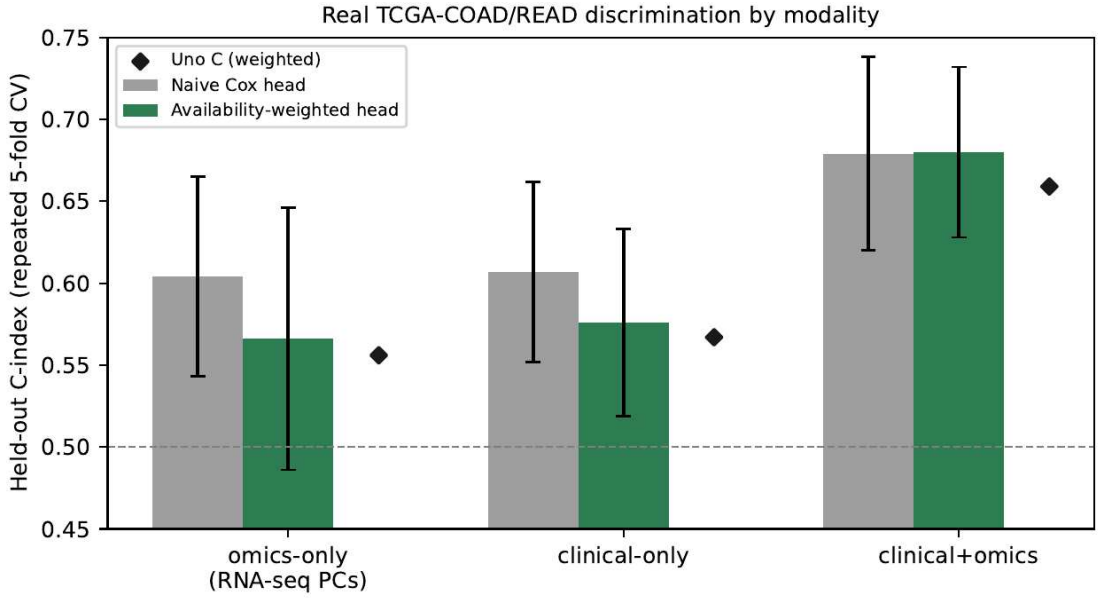
Coefficient recovery relative to the oracle. Stabilized IPW improves RMSE on METABRIC and gbsg but not on colon, GBSG2, or SUPPORT.



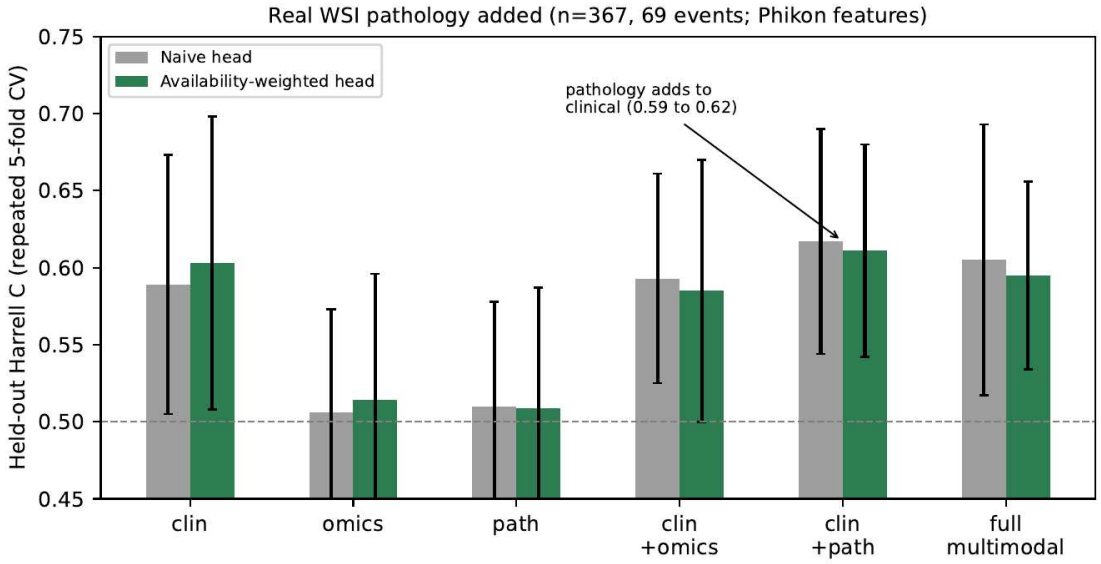
RMSE change induced by stabilized IPW (negative = improved recovery).



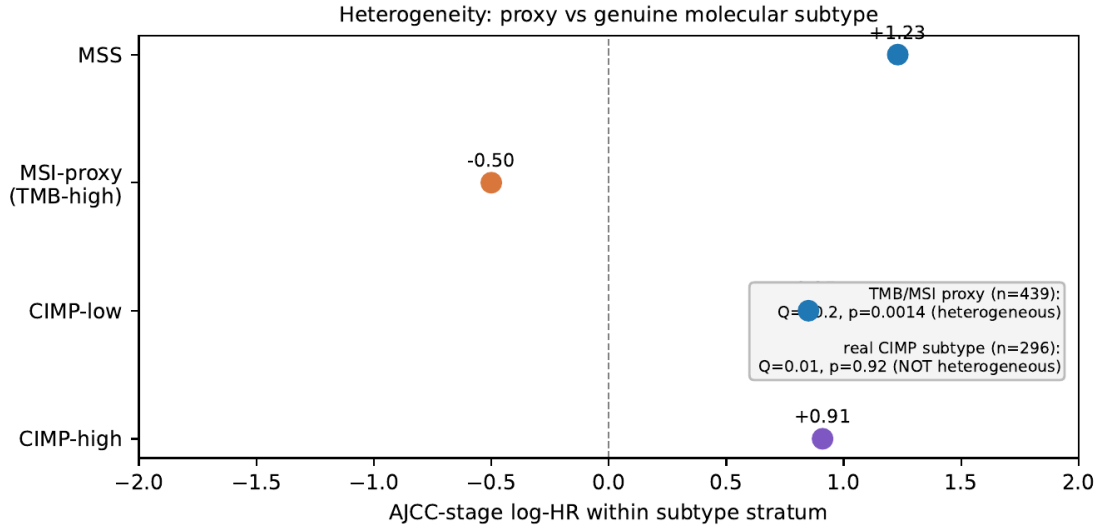
The differentiable neural stabilized-IPW Cox head improves held-out C-index over the naive neural head on all five datasets (diamonds = oracle).



Real TCGA-COAD/READ discrimination by modality ablation. Clinical+omics fusion beats single modalities; availability weighting preserves discrimination while changing the estimand.



Adding real whole-slide pathology (Phikon features): on the three-modality cohort pathology improves discrimination over clinical alone, with availability weighting preserving the ranking.



Heterogeneity of the AJCC-stage effect by molecular subtype: a TMB/MSI proxy shows significant heterogeneity (Q=10.2, p=0.001), but a genuine CIMP epigenetic subtype does not (Q=0.01, p=0.92).

Discussion

SBASURV treats sample availability as a first-class statistical object in survival learning, and the two deployments reported here—a 1,856-patient multimodal colorectal-cancer cohort and a 59,057-adult NHANES population cohort—show the same mechanism behaving

consistently across a thirty-fold range of scale and across radically different data types (tissue images and omics vs. interview and biomarker covariates). The unifying lesson is that the practical contribution is not a uniform performance gain but an *auditable procedure*. When selection is genuinely covariate-dependent, availability weighting recovers the population estimand markedly better than the naive model—demonstrably so at population scale, where it cut coefficient-recovery error by up to 62% under moderate induced selection and by 12% under the real examination-non-attendance mechanism—until weights become extreme and inverse-probability variance overwhelms the bias reduction, a regime the weight diagnostics flag in advance. When selection is near-random given covariates, as it was for the NHANES fasting subsample (propensity AUC 0.51), the correction is provably inert and leaves both estimand and discrimination untouched. A correction that changed answers under random selection would be dangerous; this negative control is itself a result.

Two deployment findings resist hype. First, availability weighting changes the estimand while leaving discrimination essentially intact—the correct behavior for a population-targeting method, and a reminder that concordance is not the metric IPW is designed to move. Second, the heterogeneity machinery delivered a consistent cross-domain caution: in colorectal cancer a proxy (TMB/MSI) subtype was significant while a genuine (CIMP) subtype was null, and in NHANES a proxy (obesity) subtype and an assay subtype (diabetes) were both significant but implied *opposite* directions of effect modification. Across tissue and population scale, effect-modification claims must be anchored to assay-grounded subtypes.

As a machine-learning contribution, SBASURV's value is its generality rather than a new network: because the head is architecture-agnostic and recovers the standard neural Cox objective when weights are uniform, any existing survival backbone—tabular MLP, sequence transformer, or attention-based multiple-instance aggregator over slide tiles—can be made population-targeting and heterogeneity-aware by swapping its loss, with no change to the encoder. The foundation-model pathway is integral to this: encoding whole-slide images with a self-supervised pathology foundation model (Phikon) and fusing those embeddings with omics and clinical data inside the weighted objective shows that selection correction can operate directly on learned high-dimensional representations, not merely on hand-specified covariates—the regime in which modern multimodal clinical AI actually operates. The same head then runs unchanged on purely tabular NHANES biomarkers, evidence that the mechanism is independent of the representation it sits on.

The pathology result is deliberately modest. A mean-pooled, label-blind foundation-model embedding contributes complementary prognostic signal on top of clinical variables but is weak in isolation at this event count, consistent with the computational-pathology literature where strong slide-level prognosis generally requires trained multiple-instance aggregation[31,32,36]. We report the conservative configuration so the multimodal claim rests on a reproducible lower bound; because the head is agnostic to the aggregator, a trained attention-MIL stream is the natural next encoder and is expected to widen the pathology contribution without any change to the weighted objective. All real results are within TCGA; generalization to prospective, multi-institution, and non-Western cohorts is

untested, and consensus molecular subtypes (CMS) could not be called from the assembled feature set[51]. The model is a research instrument for population-targeted survival learning and etiologic-heterogeneity testing, not a clinical tool; calibration, integrated Brier score[52], decision-curve analysis[53], subgroup fairness, and external replication reported following TRIPOD+AI[54] are prerequisites for any clinical reading.

Methods

Stabilized weights

With $\hat{p}_A = n^{-1}\sum_i A_i$ the stabilized weight is $\mathrm{sw}_i = \hat{p}_A/\hat{\pi}(X_i)$, clipped to $\bar{w}_i = \min\{b, \max\{a, \mathrm{sw}_i\}\}$ at the 1st/99th empirical percentiles. We report min, percentiles, mean, SD, max, and ESS $= (\sum_i \bar{w}_i)^2/\sum_i \bar{w}_i^2$ for every analysis.

Weighted Cox partial likelihood

For mini-batch B with in-batch risk set $R_B(T_i)$,

$$\mathcal{L}_{\mathrm{Cox}}(\theta) = -\sum_{i\in B} \Delta_i\, \bar{w}_i \left[f_\theta(h_i) - \log \sum_{j\in R_B(T_i)} \bar{w}_j \exp\left(f_\theta(h_j)\right)\right],$$

with the denominator implemented as $\mathrm{logsumexp}_j\{\log \bar{w}_j + f_\theta(h_j)\}$ for numerical stability; $\bar{w}_i \equiv 1$ recovers the standard neural Cox objective.

Differentiable heterogeneity head

With exposure E_i and subtype-interacted log-risk $f_{\theta,\beta}(h_i, k) = g_\theta(h_i) + \beta_k E_i$, a stratified weighted Cox objective with subtype-specific event indicators and exposure columns realizes the Lunn–McNeil construction. Under $H_0: \beta_1 = \cdots = \beta_K$, a successive-difference contrast C and robust covariance $\hat{\Sigma}_\beta$ give $Q = \left(C\hat{\beta}\right)^\top \left(C\hat{\Sigma}_\beta C^\top\right)^{-1}\left(C\hat{\beta}\right) \sim \chi^2_{K-1}$, an inferential output rather than a training penalty.

Public-benchmark protocol

We held out a 25% test set; a Cox model on the unbiased training set defined the oracle. Selection was $A \sim \mathrm{Bernoulli}\{\pi(x)\}$, $\pi(x) = \mathrm{sigmoid}(-0.2 + bz)$, with z the standardized covariate most correlated with the event. Naive and stabilized-IPW heads were trained with Adam over repeated induced-selection draws.

TCGA-CRC features and availability model

RNA-seq STAR-counts were reduced to a 16-dimensional expression-PC block; somatic mutations gave a tumor-mutation-burden feature and a hypermutation flag (MSI-H proxy)[34]. The propensity was logistic on age, sex, project, and tumor sidedness.

Pathology: up to 160 foreground tiles per slide at ~20× were embedded with Phikon[26], a publicly released iBOT[30] ViT-B/16 histopathology foundation model, and mean-pooled to a 768-dimensional slide vector reduced to 16 unsupervised PCs. CIMP was called CIMP-high when ≥3/5 methylation markers (MLH1, CDKN2A, RUNX3, SOCS1, CRABP1) were methylated[48].

NHANES cohort and population-scale availability model

We downloaded all ten continuous NHANES cycles (1999–2000 through 2017–2018) and linked each participant to the NCHS 2019 public-use Linked Mortality File[39,40]; the event is all-cause mortality and follow-up is months from interview to death or censoring. The analytic cohort is 59,057 adults (≥18 y), 9,242 deaths, 580,092 person-years. Interview-stage covariates available for all participants (age, sex, race/ethnicity, income-to-poverty ratio, smoking) define the population estimand and the availability-propensity model; examination/fasting-stage measurements (BMI, systolic blood pressure, fasting glucose, HbA1c, lipids) are present only for examined or fasting participants. The real availability indicator is measurement of the fasting biomarker panel (25,846 participants, 43.8%); a second, stronger mechanism is examination attendance (56,253 of 59,057, 95.3%). A logistic availability model yields stabilized weights $\mathrm{sw} = \hat{P}(A = 1)/\hat{\pi}(x)$ truncated at the 1st/99th percentiles[17,41]; conditioning is summarized by the Kish effective sample size and the propensity AUC. Weighted Cox models (Breslow ties, L2 penalizer 0.1, robust sandwich variance) were fit with `lifelines`[49], weights entering both the event term and the risk-set denominator. To validate the correction against a known ground truth we additionally induced covariate-dependent selection $A^* \sim \mathrm{Bernoulli}\big(\sigma(-0.2 + b \cdot \mathrm{age})\big)$ on the full cohort and swept the selection strength b. Missing covariates were median-imputed (noted in code); survey sampling weights were not combined with availability weights here and are a natural extension[55].

Use of large language models

AI-assisted software tools were used to help implement the analysis, pathology, and figure-generation code and to assist in correcting minor errors in text; all reported numbers are genuine, reproducible outputs of the released code, and the authors take full responsibility for every claim. No large language model is listed as an author.

Data availability

All datasets analyzed are publicly available: the public survival benchmarks distributed with standard survival-analysis packages; TCGA-COAD/READ data via the NIH Genomic Data Commons; and NHANES 1999–2018 with the NCHS 2019 public-use Linked Mortality File. No new data were generated.

Code availability

The complete analysis pipeline—data retrieval, cohort construction, the availability head across backbones, and figure scripts—is available from the corresponding author and regenerates every reported number from the public data.

Author contributions

Y.S.S. conceived the study, developed the method and theory, implemented all software, performed the analyses, and wrote the manuscript. S.O. provided the molecular pathological epidemiology framework, advised on the etiologic-heterogeneity analysis and interpretation, supervised the work, and revised the manuscript. Both authors approved the final version.

Competing interests

The authors declare no competing interests.

1. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 1972 34:187–202.

2. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology 2018 18:24.

3. Ching T, Zhu X, Garmire LX. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. PLoS Computational Biology 2018 14:e1006076.

4. Lee C, Zame WR, Yoon J, Schaar M van der. DeepHit: A deep learning approach to survival analysis with competing risks. Proceedings of the AAAI conference on artificial intelligence 2018 32:2314–2321.

5. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. Journal of Machine Learning Research 2019 20:1–30.

6. Wang Z, Sun J. SurvTRACE: Transformers for survival analysis with competing events. Proceedings of the 13th ACM international conference on bioinformatics, computational biology and health informatics (BCB) 2022 1–9.

7. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Annals of Applied Statistics 2008 2:841–860.

8. Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. PeerJ 2019 7:e6257.

9. Ogino S, Stampfer MJ. Lifestyle factors and microsatellite instability in colorectal cancer: The evolving field of molecular pathological epidemiology. Journal of the National Cancer Institute 2010 102:365–367.

10. Ogino S, Chan AT, Fuchs CS, Giovannucci E. Molecular pathological epidemiology of colorectal neoplasia: An emerging transdisciplinary and interdisciplinary field. Gut 2011 60:397–411.

11. Ogino S, Lochhead P, Chan AT, Nishihara R, Cho E, Wolpin BM, others. Molecular pathological epidemiology of epigenetics: Emerging integrative science to analyze environment, host, and disease. Modern Pathology 2013 26:465–484.

12. Ogino S, Ugai S, Hamada T, Ugai T. The early-onset cancer epidemics: Evidence synthesis using the prospective cohort incident-tumor biobank method. European Journal of Epidemiology 2025 40:1405–1417.

13. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 1952 47:663–685.

14. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983 70:41–55.

15. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 1994 89:846–866.

16. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004 15:615–625.

17. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. American Journal of Epidemiology 2008 168:656–664.

18. Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000 11:550–560.

19. Lin DY, Wei L-J. The robust inference for the Cox proportional hazards model. Journal of the American Statistical Association 1989 84:1074–1078.

20. Binder DA. Fitting Cox's proportional hazards models from survey data. Biometrika 1992 79:139–147.

21. Shu D, Young JG, Toh S, Wang R. Variance estimation in inverse probability weighted Cox models. Biometrics 2021 77:1101–1117.

22. Liu L, Nevo D, Nishihara R, Cao Y, Song M, Twombly TS, Chan AT, Giovannucci EL, VanderWeele TJ, Wang M, Ogino S. Utility of inverse probability weighting in molecular pathological epidemiology. European Journal of Epidemiology 2018 33:381–392.

23. Lunn M, McNeil D. Applying Cox regression to competing risks. Biometrics 1995 51:524–532.

24. Wang M, Kuchiba A, Ogino S. A meta-regression method for studying etiological heterogeneity across disease subtypes classified by multiple biomarkers. American Journal of Epidemiology 2015 182:263–270.

25. Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Song AH, others. Towards a general-purpose foundation model for computational pathology. Nature Medicine 2024 30:850–862.

26. Filiot A, Ghermi R, Olivier A, Jacob P, Fidon L, Mac Kain A, Saillard C, Schiratti J-B. Scaling self-supervised learning for histopathology with masked image modeling. medRxiv 2023 2023.07.21.23292757.

27. Xu H, Usuyama N, Bagga J, Zhang S, Rao R, Naumann T, others. A whole-slide foundation model for digital pathology from real-world data. Nature 2024 630:181–188.

28. Lu MY, Chen B, Williamson DFK, Chen RJ, Liang I, Ding T, others. A visual-language foundation model for computational pathology. Nature Medicine 2024 30:863–874.

29. Wang X, Yang S, Zhang J, Wang M, Zhang J, Yang W, Huang J, Han X. Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis 2022 81:102559.

30. Zhou J, Wei C, Wang H, Shen W, Xie C, Yuille A, Kong T. iBOT: Image BERT pre-training with online tokenizer. International conference on learning representations (ICLR) 2022.

31. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering 2021 5:555–570.

32. Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. Proceedings of the 35th international conference on machine learning (ICML) 2018 80:2127–2136.

33. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, others. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine 2019 25:1301–1309.

34. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, others. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature Medicine 2019 25:1054–1056.

35. Sirinukunwattana K, Domingo E, Richman SD, Redmond KL, Blake A, Verrill C, others. Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. Gut 2021 70:544–554.

36. Tsai P-C, Lee T-H, Kuo K-C, Su F-Y, Lee T-LM, Marostica E, others. Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. Nature Communications 2023 14:2102.

37. Chen RJ, Lu MY, Wang J, Williamson DFK, Rodig SJ, Lindeman NI, Mahmood F. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Transactions on Medical Imaging 2022 41:757–770.

38. Chen RJ, Lu MY, Williamson DFK, Chen TY, Lipkova J, Noor Z, others. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell 2022 40:865–878.e6.

39. National Center for Health Statistics. National health and nutrition examination survey (NHANES) 1999–2018. 2020.

40. National Center for Health Statistics. The 2019 public-use linked mortality files. 2022.

41. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Statistical Methods in Medical Research 2013 22:278–295.

42. Hernán MA, Robins JM. Causal inference: What if. 2020.

43. Moertel CG, Fleming TR, Macdonald JS, Haller DG, Laurie JA, Goodman PJ, others. Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. New England Journal of Medicine 1990 322:352–358.

44. Drysdale E. SurvSet: An open-source time-to-event dataset repository. 2022.

45. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 1982 247:2543–2546.

46. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012 487:330–337.

47. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine 2011 30:1105–1117.

48. Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, Faasse MA, others. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nature Genetics 2006 38:787–793.

49. Davidson-Pilon C. Lifelines: Survival analysis in Python. Journal of Open Source Software 2019 4:1317.

50. Menke A, Casagrande S, Geiss L, Cowie CC. Prevalence of and trends in diabetes among adults in the United States, 1988–2012. JAMA 2015 314:1021–1029.

51. Guinney J, Dienstmann R, Wang X, Reyniès A de, Schlicker A, Soneson C, others. The consensus molecular subtypes of colorectal cancer. Nature Medicine 2015 21:1350–1356.

52. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine 1999 18:2529–2545.

53. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making 2006 26:565–574.

54. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, others. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 385:e078378.

55. Lumley T. Analysis of complex survey samples. Journal of Statistical Software 2004 9:1–19.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SBASURVPresentation.pptx