Guarantees That Survive a Missing Scan: Modality-Conditional Conformal Prediction for Multimodal Medical Diagnosis

Aaron Ajit[0009−0001−5768−2837]

Independent Researcher, Irving, TX, USA
aaronaajit@gmail.com

Abstract. Multimodal models that combine medical images with tabular clinical data are routinely deployed on *incomplete* inputs: a patient may arrive without a scan, or without labs. Conformal prediction (CP) is the standard way to attach a distribution-free coverage guarantee to such models, but its guarantee holds only under exchangeability of the calibration and test data. We show that calibrating on modality-complete data breaks this assumption for test cases missing an entire modality, and standard (marginal) split CP silently loses *conditional* validity on the modality-incomplete subgroup. We further show that this failure has *two faces*: on one dataset marginal CP remains valid but collapses to uninformative prediction sets, while on another it retains informativeness only by *violating* coverage (dropping to 0.73 against a 0.90 target in the longer-trained regime). We introduce Modality-Conditional Conformal prediction (MC^2), which stratifies calibration by the modality-availability pattern (a Mondrian taxonomy) and thereby restores valid per-pattern coverage with finite-sample guarantees, requiring no retraining of the predictor. Across two public image+tabular datasets (CBIS-DDSM mammography; OL3I cardiac CT) and a controlled multi-class study, MC^2 restores the proper validity–informativeness trade-off in both failure regimes. Through a controlled prevalence-swap experiment we further identify *class prevalence*, not modality redundancy, as the driver of which face appears, and give a concrete threshold-shift mechanism.

Keywords: Conformal prediction · Uncertainty quantification · Multimodal learning · Missing modalities · Trustworthy AI.

1 Introduction

Clinical decision support increasingly fuses imaging with structured tabular data (labs, demographics, risk factors). Two facts about real deployment collide. First, modalities go missing: not every patient has a current scan, and not every scan is paired with complete labs, so models are evaluated on incomplete inputs constantly. Second, clinicians need a calibrated notion of trust, not just a point prediction. Conformal prediction (CP) addresses the second need: it wraps any predictor and returns a set of labels guaranteed to contain the truth with probability at least $1 - \alpha$, distribution-free and finite-sample.

The guarantee, however, rests on *exchangeability* between calibration and test data—not broken by modality absence *per se*, but by a mismatch between calibration- and test-time missingness: if calibration is performed on modality-complete patients but a test patient is missing the image, the non-conformity scores are no longer exchangeable, and the $1 - \alpha$ guarantee quietly voids—precisely when the input is degraded and trust matters most. The reported coverage still "says 0.90"; the realized coverage on the incomplete subgroup is different.

Our contributions are:

1. We formalize and empirically demonstrate the loss of *conditional* validity of split CP under whole-modality absence, and show it has *two empirical faces*: *valid-but-uninformative* (sets degenerate to the full label space) and *informative-but-invalid* (coverage drops below target).
2. We introduce $\mathbf{MC}^2$, modality-conditional conformal prediction: Mondrian calibration with the grouping function set to the modality-availability pattern, restoring per-pattern coverage with finite-sample validity and no change to the trained predictor.
3. We evaluate on two public image+tabular medical datasets and a controlled multi-class study (30 seeds, 95% CIs), characterize score-dependence and robustness, and show by a controlled prevalence swap that class prevalence, not modality redundancy, governs which face appears.

2 Related Work

Conformal prediction in medical imaging. CP [13, 1] has been applied to disease severity rating, fairness-aware predictive sets [3], and medical vision–language models, and studied under distribution shift across sites—all assuming complete, identically distributed inputs, not whole-modality absence. **Missing-modality fusion.** A large body of work makes the *predictor* robust to missing modalities via modality dropout, knowledge distillation, contrastive alignment, and learnable mask tokens [5, 14]; these improve accuracy/AUROC under missingness but give no statistical guarantee on the resulting uncertainty. The one prior CP-with-incomplete-modalities use we know of, Any2Any [7], targets multimodal *retrieval*, aligning similarity scores—a different task and role for CP than per-pattern coverage in classification. **Group-conditional CP.** Mondrian conformal prediction [13, 4] guarantees coverage within each cell of a partition defined by a grouping function over covariates; class-conditional variants are the standard fix for imbalance-induced miscoverage. Full conditional coverage is impossible distribution-free [2]. Mondrian CP is an established tool; our contribution is instantiating its grouping function as the modality-availability pattern—to our knowledge not previously applied to missing-modality multimodal medical diagnosis—and showing empirically why it is necessary (Sect. 5.1) and where fragile (Sect. 5.2). Prior work makes the *prediction* robust; we make the *guarantee* robust. **Relation to weighted CP.** Weighted CP [12] handles covariate

shift via density-ratio reweighting; here the shift variable is observed and discrete, so Mondrian conditioning gives an exact per-pattern guarantee with no ratio estimation—weighted CP becomes the natural route if missingness is informative (Sect. 6).

3 Method

3.1 Problem setup

Each patient i has an image x_i^{img}, a tabular vector x_i^{tab}, a label $y_i \in \{1, \ldots, K\}$, and an availability mask $m_i \in \{0,1\}^2$ indicating which modalities are present. Define the modality-availability pattern $g(x_i) = m_i \in \mathcal{M} = \{(1,1),(1,0),(0,1)\}$ (both / image-only / tabular-only). We are given a predictor f producing a probability vector $p = f(x)$ under any subset of modalities.

3.2 A missing-modality-capable predictor

The CP layer needs a predictor that outputs a calibrated-in-rank softmax under *any* modality subset; this part is standard and not our contribution. We use a lightweight fusion MLP over a *frozen* image-encoder embedding and a small tabular encoder, with a learnable mask token for any absent modality and modality dropout during training so f observes incomplete inputs. Freezing the image encoder keeps training feasible on commodity hardware and isolates our contribution to the calibration layer.

3.3 Modality-Conditional Conformal prediction (MC^2)

Let $s(x,y)$ be a non-conformity score on $p = f(x)$; we consider LAC [11], $s = 1 - p_y$, and APS (adaptive prediction sets) [10]. Given a calibration set of size n, *marginal* split CP computes a single threshold $\hat{q} = \lceil (n+1)(1-\alpha) \rceil / n$-quantile of $\{s(x_i, y_i)\}$ and returns $C(x) = \{y : s(x,y) \le \hat{q}\}$, with the marginal guarantee $\mathbb{P}(Y \in C(X)) \ge 1 - \alpha$ under exchangeability.

MC^2 instead calibrates one threshold *per modality pattern*. For each $m \in \mathcal{M}$,

$$\hat{q}_m \;=\; \Big\lceil (n_m+1)(1-\alpha) \Big\rceil / n_m\text{-quantile of } \{\, s(x_i,y_i) : g(x_i) = m \,\}, \qquad (1)$$

i.e. the smallest score whose empirical rank is at least $\lceil (n_m+1)(1-\alpha) \rceil$ out of n_m, applied *within* pattern m rather than pooled (n_m = calibration points with $g(x_i) = m$); with small n_m this ceiling can round up to the maximum score, giving a vacuous threshold, which motivates the τ fallback below. The test point uses $\hat{q}_{g(x)}$. By the Mondrian construction this yields the *conditional* guarantee

$$\mathbb{P}\big(Y \in C(X) \,\big|\, g(X) = m\big) \;\ge\; 1 - \alpha \qquad \text{for every } m \in \mathcal{M}, \qquad (2)$$

under exchangeability *within* each pattern—requiring calibration and test scores to be exchangeable only *conditional on the missingness pattern being the same*,

a strictly weaker and realistic assumption than global exchangeability. We populate the incomplete patterns by applying modality dropout to the calibration split, so within a pattern the calibration and test missingness mechanisms coincide by construction (Sect. 6 discusses informative missingness). Groups with fewer than $\tau = 20$ calibration points fall back to the pooled threshold (reported, not hidden; behavior characterized in Sect. 5.3). No headline result triggers it: image-missing calibration sizes range $n \approx 40$–400 across the CBIS/OL3I headlines and prevalence-swap arms, all above τ. Per-pattern storage is $|\mathcal{M}|$ scalars; an unseen pattern uses the pooled threshold ($n_m = 0$) until the next recalibration.

4 Experimental Setup

Datasets. (i) *CBIS-DDSM* [6]: mammography ROIs with tabular metadata (breast density, abnormality type, subtlety, view); a binary benign/malignant label (prevalence 40.8%; 3,567 cases, 1,566 patients). We exclude the BI-RADS assessment field to avoid label leakage. (ii) *OL3I* (Stanford AIMI): a single L3 CT slice per patient with tabular clinical risk factors (age, sex, BMI, smoking, glucose, systolic/diastolic BP); binary 1-year ischemic heart disease (prevalence 4.4%; 8,139 patients) [15]. We exclude the ICD/CPT/ATC encounter block (unauditable pre-index lookback) and the Charlson index (overlaps the outcome). **Encoder.** A frozen DINOv2 ViT-S/14 [8] (384-d embeddings), extracted once; CT slices are windowed and replicated to three channels. **Protocol.** 50/25/25 patient-grouped, label-stratified train/calibration/test split; OL3I trained with class-weighted loss. Mean ±95% CI over 30 random splits (redrawing both the partition and missingness masks); target $\alpha = 0.1$; calibration patterns populated at a modality-dropout rate of 0.20; two training regimes, early-stopped (CBIS ∼25, OL3I ∼20 epochs) and longer-trained (∼200 epochs). Besides marginal split CP and MC^2, Figs. 2 and 3 include an *imputation* baseline (defined in the Fig. 2 caption). Code and full results: see footnote.[1]

5 Results

5.1 The two faces of the failure (LAC)

Table 1 reports the image-missing (`tab_only`) subgroup. On both datasets, marginal CP fails to be simultaneously valid and informative, but it fails in opposite ways.

CBIS — valid but uninformative. Marginal CP *satisfies* coverage on the image-missing group (1.000 at 200 epochs) only by emitting the full two-class set for every case (set size 2.00, actionable rate 0.000): the guarantee is met vacuously. MC^2 holds coverage near target (0.907) while returning a confident

[1] `https://github.com/aaronjji/Modality-Conditional-Conformal-Prediction-for-Multimodal-Medical-Diagnosis`

Table 1. Image-missing (`tab_only`) subgroup, LAC score, target coverage 0.90. "Act." is the fraction of confident single-class (actionable) predictions; "Cov." is empirical coverage. Marginal CP fails differently on the two datasets; MC^2 restores valid coverage in both.

Dataset (regime)	Method	Cov.	Act.	Set size
CBIS-DDSM (200 ep)	Marginal	1.000	0.000	2.00
	MC^2	0.907	0.245	1.76
CBIS-DDSM (25 ep)	Marginal	0.958	0.135	1.87
	MC^2	0.910	0.263	1.74
OL3I (200 ep)	Marginal	**0.728**	0.892	1.10
	MC^2	0.908	0.535	1.47
OL3I (20 ep)	Marginal	0.889	0.534	1.47
	MC^2	0.905	0.502	1.50

single-class call on 24.5% of cases—roughly doubling actionability at 25 epochs (0.263 vs. 0.135) and rescuing it from zero when the model is longer-trained.

OL3I — informative but invalid. Here marginal CP *violates* coverage on the image-missing group, falling to 0.728 (a 17-point shortfall; 95% CI well clear of target) at 200 epochs. Its high apparent actionability (0.892) is illusory: the truth is missed 27% of the time—driven by over-calling the positive class at roughly 6× its true prevalence. MC^2 restores valid coverage (0.908) at the honest cost of fewer, trustworthy confident calls. The violation emerges with training: mild at 20 epochs (0.889), severe at 200 (0.728); we report both and do not headline the overfit figure as typical.

The other patterns, for completeness. We headline `tab_only` because that is where marginal CP fails; on the `both` and `img_only` patterns marginal coverage stays at or above target on OL3I (0.935, 0.936) and within five points on CBIS (0.884, 0.850)—nowhere near OL3I's 17-point `tab_only` shortfall (full results: code repository).

In both cases the unifying statement holds: *marginal CP cannot be both valid and informative on the modality-incomplete subgroup; MC^2 restores the proper trade-off* (per-rate curves: Fig. 2).

The coverage face, made explicit. Because both datasets are binary, their coverage face is necessarily muted (a binary set degenerates rather than visibly under-covering). A controlled multi-class synthetic study makes it explicit (Fig. 1): with $K = 6$ and an image-missing test subgroup, marginal CP under-covers severely and flatly—0.455 ± 0.012 against a 0.90 target, independent of the test-time missingness rate since miscalibration is fixed at calibration time—while MC^2 holds 0.90 throughout: the multi-class end of a label-granularity continuum whose binary end is Table 1's degeneracy/violation behavior.

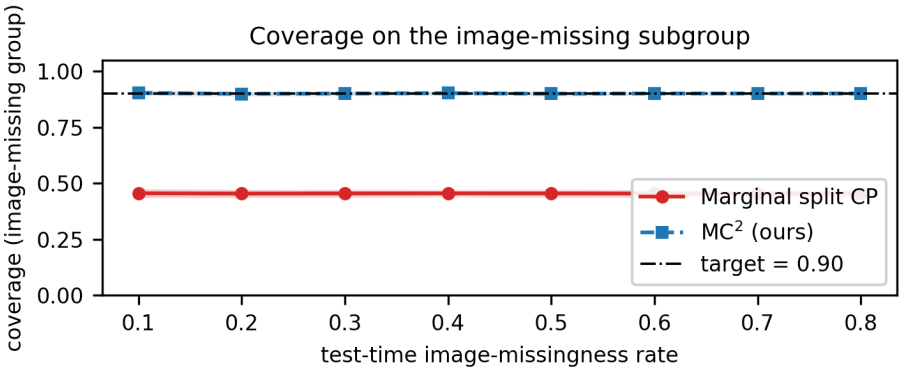

Fig. 1. Controlled multi-class synthetic study ($K = 6$, 30 seeds, 95% CI). On the image-missing subgroup, marginal split CP under-covers by ≈ 45 points (0.455 vs. the 0.90 target), flat across the missingness rate; MC^2 holds target. This is the coverage face the binary clinical datasets can only express indirectly.

5.2 Score dependence

The effect is score-dependent, and we scope our primary claim to LAC. Under LAC, MC^2's advantage is positive and far from zero in both regimes (+0.13 and +0.25 actionability on CBIS). Under APS—an adaptive score inflating set size for uncertain inputs—the comparison is regime-dependent and, on small per-pattern groups, can reverse (CBIS early-stopped favors marginal CP, longer-trained does not). A deconfounded synthetic study at matched small n does not reproduce the reversal, so it is not a simple finite-sample artifact; we report it without claiming a mechanism (Fig. 3, bottom)—consistent with findings that Mondrian calibration on thin cells can overreact to local score idiosyncrasies [9].

5.3 Ablations and robustness

Coverage level. Across $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$ MC^2 tracks target coverage while marginal CP stays near-degenerate in the overfit regime. **Group size and fallback.** As τ (`min_group_size`) is swept past a group's calibration size, the transition is clean: while τ is below that size MC^2 operates per-pattern; once τ exceeds it, MC^2 collapses exactly onto marginal CP's reference values, as designed. **Within-patient correlation.** One case per patient leaves the result intact (CBIS 200-epoch: marginal 1.000/0.000, MC^2 0.923/0.161), not an artifact of repeated cases. **Calibration composition.** Raising the calibration dropout rate populates the incomplete patterns and improves MC^2 actionability.

5.4 Why do the two datasets fail differently? Prevalence, not redundancy

A natural first hypothesis is modality *redundancy*: if the tabular modality is redundant with the image, dropping the image is cushioned. Two observations say

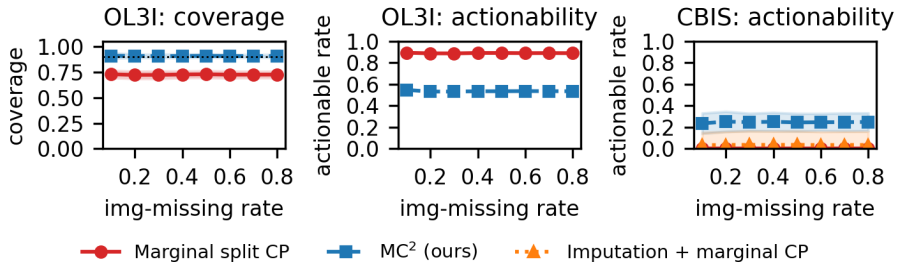

Fig. 2. Left two panels (OL3I, coverage face): marginal CP under-covers the image-missing group while MC^2 holds target. Right panel (CBIS, informativeness face): MC^2 recovers actionability the marginal baseline loses; *Imputation + marginal CP* (mean-impute the missing embedding, then one marginal threshold) recovers less and carries no per-pattern guarantee. Curves are mean over 30 seeds.

Table 2. Prevalence-swap (image-missing subgroup, LAC, 200 epochs, 30 seeds). Changing only the positive rate moves the failure face. Bold: coverage violation. † small calibration group, reported not hidden.

Dataset	Prevalence	Marginal cov.	MC^2 cov.
CBIS	40.8% (orig.)	1.000	0.907
CBIS	4.4% (rare)	**0.747**	0.916
OL3I	4.4% (orig.)	**0.728**	0.908
OL3I	40% (rebal.)†	1.000	0.945

otherwise. First, the observational measure is confounded and does not cleanly separate the datasets: mean cross-validated R^2 of predicting tabular features from the image is, if anything, *higher* for CBIS (0.33) than OL3I (0.28), and in both cases is dominated by near-tautologically image-readable features, while OL3I's *predictive* risk factors are genuinely complementary (per-feature values: code repository); we do not rest the explanation on this comparison. Second, and decisively, the prevalence swap below holds each dataset's modality structure—hence redundancy—fixed while changing only the positive rate, yet flips the failure face: no redundancy-based account can explain that.

The operative variable is prevalence. We isolated it with a controlled swap, holding everything else fixed and changing only the positive rate by *down-sampling* (never duplicating) patients (Table 2), at the longer-trained regime where the effect is sharpest. Sub-sampling CBIS to 4.4% converts its valid-but-degenerate behavior into a genuine coverage *violation* (0.747 $[0.722, 0.773]$, matching OL3I's signature); rebalancing OL3I toward 40% converts its violation into the degenerate-but-valid behavior (1.000, near-zero actionability, matching CBIS). The well-powered CBIS arm ($n_{\text{cal}} = 91$) carries the conclusion; the OL3I arm is confirmatory at small n (≈ 40; flagged).

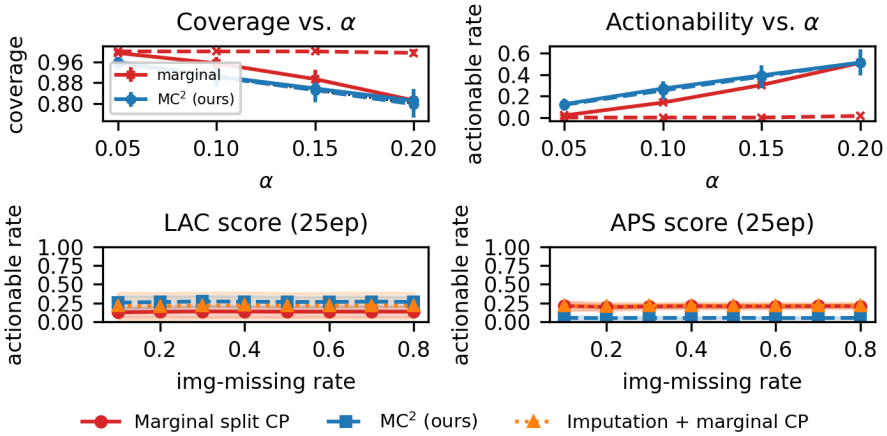

Fig. 3. Top: α sweep (solid: 25 epochs; dashed: 200 epochs)—MC^2 tracks target coverage. Bottom: LAC vs. APS—the actionability advantage is LAC-specific; the imputation baseline (*Imputation + marginal CP*) is shown for reference.

Mechanism. The split reflects how a marginal LAC threshold behaves under imbalance once a subgroup's score distribution diverges from the pool. At low prevalence the modality-complete calibration scores shrink toward zero with training (the model confidently nails the majority class), dragging the pooled threshold *below* the image-missing group's score ceiling and producing undercoverage; at higher prevalence, harder classification eventually yields a few confidently wrong complete-modality points, *inflating* the pooled threshold past that ceiling, yielding conservative, degenerate sets. The missing modality *creates* the divergent subgroup; prevalence sets the mismatch direction; the effect is markedly weaker early-stopped, where swapped cells' marginal coverages cluster near target—a prevalence×training-duration interaction. MC^2 removes the dependence in all configurations by calibrating the image-missing group on its own scores; the mechanism is consistent-with rather than proven here.

6 Limitations

The image-missing calibration group is modest on CBIS ($n \approx 163$, SE$\approx$ 2.3pp); the headline relies on a frozen encoder, so fine-tuning could shift numbers. Our datasets are binary; the multi-class face is synthetic-only. The APS regime-dependence remains open, and our prevalence mechanism (Sect. 5.4) rests on a small OL3I group. Exchangeability is assumed at the case level within a patient-grouped split (dedup mitigates, not eliminates, within-patient correlation) and conditionally at random within each pattern—cells built by dropout, so calibration/test are exchangeable by construction, but *informative* missingness could break this; the guarantee is unconditional on class too (OL3I image-missing positive/negative coverage: 0.650/0.726 marginal, 0.826/0.910 MC^2).

7 Conclusion

A test-time missing modality breaks the exchangeability CP depends on when calibration is modality-complete: standard CP becomes uninformative or invalid on exactly the degraded patients. MC^2 restores finite-sample, per-pattern coverage with no predictor change—a simple, post-hoc, cheap fix for a failure ubiquitous in multimodal clinical deployment (open directions: Sect. 6).

Disclosure of Interests. The author has no competing interests to declare.

Generative AI Usage. The author used a large language model to help draft manuscript text and to write and edit analysis code, all of which the author reviewed and verified. The tool was not used to generate, alter, or select reported results; the author takes full responsibility for the accuracy and integrity of the paper.

References

1. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021)
2. Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J.: The limits of distribution-free conditional predictive inference. Information and Inference: A Journal of the IMA **10**(2), 455–482 (2021)
3. Cresswell, J.C., Kumar, B., Sui, Y., Belbahri, M.: Conformal prediction sets can cause disparate impact. In: International Conference on Learning Representations (ICLR) (2025), arXiv:2410.01888
4. Gibbs, I., Cherian, J.J., Candès, E.J.: Conformal prediction with conditional guarantees. Journal of the Royal Statistical Society Series B: Statistical Methodology **87**(4), 1100–1126 (2025)
5. Kim, D., Kim, T.: Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In: European Conference on Computer Vision (ECCV) (2024), arXiv:2407.12616
6. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific Data **4**, 170177 (2017)
7. Li, P.H., Yang, Y., Omama, M., Chinchali, S., Topcu, U.: Any2any: Incomplete multimodal retrieval with conformal prediction. arXiv preprint arXiv:2411.10513 (2024)
8. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (TMLR) (2024), arXiv:2304.07193
9. Rafe, A., Das, S.: Socio-conformal calibration in complex survey data: Marginal validity is not enough for subgroup reliability. arXiv preprint arXiv:2605.05562 (2026)

10. Romano, Y., Sesia, M., Candès, E.J.: Classification with valid and adaptive coverage. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 3581–3591 (2020)
11. Sadinle, M., Lei, J., Wasserman, L.: Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association **114**(525), 223–234 (2019)
12. Tibshirani, R.J., Barber, R.F., Candès, E.J., Ramdas, A.: Conformal prediction under covariate shift. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 32 (2019)
13. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer (2005)
14. Xu, Y., Zhou, F., Zhao, C., Wang, Y., Yang, C., Chen, H.: Distilled prompt learning for incomplete multimodal survival prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025), arXiv:2503.01653
15. Zambrano Chaves, J.M., Wentland, A.L., Desai, A.D., Banerjee, I., Kaur, G., Correa, R., Boutin, R.D., Maron, D.J., Rodriguez, F., Sandhu, A.T., Rubin, D., Chaudhari, A.S., Patel, B.N.: Opportunistic assessment of ischemic heart disease risk using abdominopelvic computed tomography and medical record data: A multimodal explainable artificial intelligence approach. Scientific Reports **13**, 21034 (2023), oL3I dataset: `https://stanfordaimi.azurewebsites.net/datasets/3263e34a-252e-460f-8f63-d585a9bfecfc`