CARE-FUSE: RELIABILITY-AWARE MULTIMODAL FUSION UNDER MISSING AND CONFLICTING CLINICAL EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal fusion must remain effective when evidence is missing or expert predictions conflict. We introduce CARE-Fuse, a task-adaptive late-fusion framework that incorporates evidence availability and cross-expert disagreement into validation-trained gates. A reliability gate is trained with label-independent validation perturbations to improve robustness to missing and conflicting inputs. We evaluate CARE-Fuse on clinical prediction tasks by combining first-24-hour vital signs and laboratory features, temporal representations, report embeddings, and language-model scores. On a locked test cohort of 14,637 MIMIC ICU stays, CARE-Fuse achieves AUROC/AUPRC of 0.8674/0.5344 for in-hospital mortality and 0.8310/0.6878 for stays longer than three days. Relative to the strongest single expert, paired bootstrap analysis with 5,000 draws yields gains of 0.0428 AUROC and 0.0671 AUPRC for mortality, and 0.0386 AUROC and 0.0960 AUPRC for prolonged stay; all 95% confidence intervals exclude zero. Under missing temporal or report evidence and conflicting language-model outputs, the perturbation-trained reliability gate reduces degradation compared with an unaugmented gate, while a clinical-only stack remains competitive on clean data. Calibration and selective-prediction analyses further characterize evidence coverage, expert disagreement, and case-review behavior. These findings support perturbation-trained fusion with explicit availability and disagreement features as a practical approach to robustness in multimodal clinical prediction.

Keywords: Multimodal learning, missing-modality robustness, clinical risk prediction, reliability-aware fusion, expert disagreement, selective prediction

1 INTRODUCTION

A predictive system may use the same model output to order cases, display a probability, and select records for further processing. Ordering depends on discrimination; probability interpretation depends on calibration; budgeted selection depends on which errors enter the selected set. In multimodal systems, these operations also depend on evidence availability. A single performance ranking leaves their operational differences unresolved.

One concrete mismatch arises when balanced accuracy selects a classifier threshold but total-error capture evaluates a downstream ranker. These criteria assign different relative weights to false negatives and false positives. Carrying the threshold-selection weights into ranking evaluation can favor a different policy even when predictions remain fixed. In the archived prolonged-stay task, increasing the FN/FP cost ratio from one to two reverses the preferred ranking at a 10% selection budget. Such comparisons help specify the prediction and case-queue components used in human–AI systems (Amershi et al., 2019; Bansal et al., 2021; Sivaraman et al., 2025).

CARE-Fuse examines frozen multimodal clinical outputs through four fusion variants: a clinical stack, simple five-score fusion (S), engineered fusion with availability and disagreement features (U), and a perturbation-trained variant (R). Three fixed rankings are then applied to U. We contribute:

- An error-objective analysis connecting validation threshold selection to fixed-budget ranking. Validation balanced accuracy implies FN/FP weights of approximately 7.66 for mortality and

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.

2.17 for LOS, above the respective 10% selection crossovers of 6.118 and 1.571. Selected and retained confusion counts explain the policy reversal.

- A comparator analysis showing that simple five-score fusion narrows the apparent robustness benefit under joint temporal/report-output removal: at 50% removal, mean LOS AUROC is 0.8253 for R, 0.8232 for S, and 0.7240 for U. R has the highest minimum condition mean across the tested stresses.
- Coverage and output diagnostics distinguishing referenced from readable images, and continuous ranking changes from hard-decision changes under image replacement.

2 Related Work

Incomplete clinical fusion is addressed by MedFuse, MultiModN, DrFuse, and confidence-guided MedPatch (Hayat et al., 2022; Swamy et al., 2023; Yao et al., 2024; Al Jorf & Shamout, 2025). CARE-Fuse fixes the archived scalar-score interface to examine comparator sensitivity, rather than learning new modality representations. Calibration and selective classification distinguish score interpretation from ranking and rejection (Guo et al., 2017; Geifman & El-Yaniv, 2017; Mozannar & Sontag, 2020). Cost-sensitive rejection and evaluation studies show that objectives can change method rankings (Franc et al., 2023; Swaminathan et al., 2024; Traub et al., 2024; Narasimhan et al., 2024); prior empirical work also reports FP/FN cost-dependent reversals (Gupta & Liu, 2026). Our contribution is a fixed-classifier decomposition of these choices in two clinical outcomes, including crossovers relative to validation threshold-selection weights. This supplies computational evidence for inspectable evaluation tools such as Zeno, Tempo, and Orbit (Cabrera et al., 2023; Sivaraman et al., 2025; Yang et al., 2025). Appendix D develops the broader comparison with clinical benchmarks, uncertainty evaluation, and reproducibility literature.

3 Data and Evaluation Design

We analyze frozen results derived from MIMIC-IV v2.2, MIMIC-IV-Note v2.2, and MIMIC-CXR-JPG v2.1.0 (Johnson et al., 2023a;b; 2024). The archived reconstruction partitions 73,181 ICU stays into 51,226 training, 7,318 validation, and 14,637 test records. Outcomes are in-hospital mortality and ICU length of stay greater than three days (LOS). Inputs nominally describe the first 24 hours, while outcomes describe the whole stay.

The archive contains cohort summaries, model and protocol metrics, stress aggregates, selection summaries, code, and configurations. Its main split uses reindexed stay rows; the encoder reconstruction uses stored split labels assigned before sorting. Modality cutoffs differ, and cohort construction does not require survival or continued ICU occupancy through 24 hours. Missing patient–stay mappings and event/availability timestamps prevent checks of patient overlap or common-landmark eligibility. Exact test QA–stay agreement and zero validation/test label mismatches do not resolve those limitations. Appendix C gives extraction rules and new cohort-checking tools evaluated on synthetic fixtures.

Three questions organize the analysis: **RQ1**, how coverage and output protocols affect score and decision interpretation; **RQ2**, how comparator choice changes clean discrimination, calibration, and stress conclusions; and **RQ3**, how fixed-budget rankings change error composition and policy preference across error costs. RQ2 compares clinical/S/U/R; RQ3 uses U only (Figure 1). Initial-fusion comparisons against individual experts are a separate configuration (Appendix A).

4 Methods

4.1 Expert Scores and Fusion

The five scalar inputs are structured vital-sign/demographic/admission scores p_v, laboratory-summary scores p_l, GRU-representation scores p_g, RadBERT-representation scores p_r, and a Luna auxiliary channel p_a. Clinical experts use median imputation, missing-cell indicators, standardization, and class-balanced logistic classifiers; large embeddings use a log-loss SGD classifier. Base experts are fitted on training. Images reach fusion through the auxiliary channel; Sol is a separate comparator.

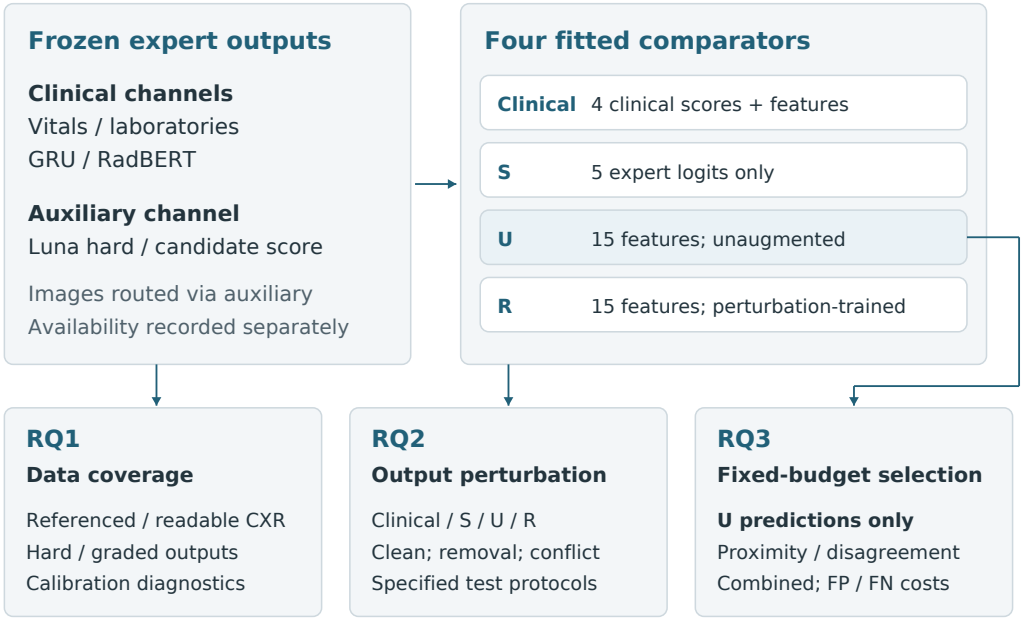


Figure 1: The three analysis paths connect input coverage, fusion response, and error allocation. RQ2 compares the clinical stack, scores-only S, CARE-Fuse-U, and CARE-Fuse-R; RQ3 uses CARE-Fuse-U.

Unavailable GRU/report branches use prevalence among their available training records and clear masks m_g, m_r. The auxiliary channel uses a finite candidate score when available, otherwise parsed No/Yes mapped to $\epsilon/1-\epsilon$, with $\epsilon = 10^{-5}$. All scores are clipped to this interval. Mask m_s denotes candidate-score availability; m_i denotes image routing, not verified readability. With $\ell(p) = \log[p/(1-p)]$ and $\mathbf{p} = (p_v, p_l, p_g, p_r, p_a)$, U and R use

$$\mathbf{x} = \big[\ell(p_v), \ell(p_l), \ell(p_g), m_g, \ell(p_r), m_r, \ell(p_a), m_i, \\ \text{mean}(\mathbf{p}), \text{sd}(\mathbf{p}), \text{range}(\mathbf{p}), \ell(p_a)m_i, \ell(p_a)m_s, |p_a - p_g|, |p_a - p_l|\big], \tag{1}$$

$$s_t = \sigma\big(b_t + \mathbf{w}_t^{\mathsf{T}} \,\text{Standardize}(\mathbf{x})\big). \tag{2}$$

Task-specific coefficients are fixed after fitting; interactions express dependence on evidence state. S uses only five logits. The clinical stack excludes the auxiliary score and uses four clinical logits, GRU/report and image-routing flags, and mean, population standard deviation, and range over its four scores.

Clinical/S/U fit standardized, class-balanced, L2 logistic gates on validation. Five-fold stratified cross-validation selects $C \in \{0.001, 0.01, 0.1, 1, 10\}$ by pooled out-of-fold AUROC. Thresholds maximize out-of-fold balanced accuracy over unique score quantiles at 193 positions between 0.02 and 0.98; gates are refitted on all validation records. All variants expose raw sigmoid scores without post-hoc calibration.

4.2 Perturbation Training and Evaluation

R augments validation with GRU removal, report removal, joint removal, and auxiliary conflict. Each family selects 10%, 30%, or 50% of all records using three augmentation seeds. The clean block plus 36 augmented blocks yields 37 variants per record, including unchanged copies. Removal substitutes full-training-cohort prevalence and clears the affected mask; report removal also clears image routing while retaining the archived auxiliary score. Conflict sets $p_a = 1 - \text{mean}(p_v, p_l, p_g, p_r)$, preserving candidate-score availability. Vitals and laboratory scores remain intact. Selection and alteration are label-independent; gate fitting is supervised.

All copies of each source record remain in the same cross-validation fold. R selects regularization and threshold on clean held-out validation blocks before fitting all augmented blocks. U and R have different fitting seeds and separately selected hyperparameters. Test stress uses the same four families and three rates, with five draws from seeds disjoint from augmentation. We report draw means, each model's minimum across the 12 nonzero conditions, and its maximum gap to the condition-specific best mean. S is aligned to R by outcome, condition, rate, and seed, checking shared U/clinical cells. Available paired-bootstrap R–U and R–clinical contrasts concern only the first draw; S comparisons are descriptive.

4.3 Fixed-Budget Selection and Error Costs

For refitted U score s and threshold τ, define

$$P = 1 - \text{clip}\left(|s - \tau| / \max(\tau, 1 - \tau, \epsilon), 0, 1\right), \tag{3}$$

$$D = \text{sd}(\mathbf{p}), \qquad M = (1 - m_g) + (1 - m_r) + (1 - m_s) + (1 - m_i), \tag{4}$$

$$\rho = [F_P(P) + F_D(D) + F_M(M)]/3. \tag{5}$$

The empirical CDFs F_P, F_D, F_M use validation; P uses in-sample scores of the refitted gate. For $b \in \{0.05, 0.10, 0.20, 0.30\}$, combined, proximity, and disagreement select $m = \text{round}(bN)$ records by $\rho, F_P(P), F_D(D)$. CDFs are right-inclusive and the archive reverses default `numpy.argsort` without a stable tie key. Predictions remain fixed. Error capture and FN capture divide selected errors by all errors and selected FN by all FN, respectively. Enrichment divides error capture by the actual selected fraction.

Aggregate capture fractions recover selected confusion counts; subtraction yields retained counts. All 76 floating-to-integer checks agree within 10^{-12}. With FN/FP cost ratio $\lambda > 0$, retained weighted error and a pairwise crossover are

$$L_{\text{ret}}(\pi, b; \lambda) = \frac{\text{FP} - \text{FP}_{\text{sel}}(\pi, b) + \lambda[\text{FN} - \text{FN}_{\text{sel}}(\pi, b)]}{N - m}, \tag{6}$$

$$\lambda^* = \frac{\text{FP}_{\text{sel,A}} - \text{FP}_{\text{sel,B}}}{\text{FN}_{\text{sel,B}} - \text{FN}_{\text{sel,A}}}. \tag{7}$$

At a fixed budget, minimizing retained cost maximizes selected FP plus λ times selected FN. A positive ratio defines a positive-cost crossover; equal FN counts give parallel or coincident lines. Exact rational intersections and lower envelopes compare the three fixed policies (Drummond & Holte, 2006). Costs are sensitivity parameters, not estimated clinical utilities. Selected cases are excluded, not assumed corrected. Appendix F.5 details count reconstruction and risk denominators.

4.4 Output Diagnostics and Statistics

Image replacement holds model, text, and prompt fixed within the readable-image subset; missing-image records receive identical placeholders. Hard decisions and candidate scores are separate channels. LLaVA-Med uses next-token Yes/No softmax mass and threshold 0.5. API scores normalize the first generated position exposing both labels; other outputs are parsed decisions. Appendix F specifies extraction and protocols.

We report AUROC, non-interpolated average precision (AUPRC for main fusion), hard-decision metrics, Brier score, and 15-bin equal-width ECE. Test-prevalence $\pi(1 - \pi)$ supplies a descriptive Brier reference. Outcome-stratified, paired record bootstraps use 2,000 draws for first-draw stress and image comparisons; initial fusion uses 5,000 (Appendix A). Nominal 95% intervals omit within-patient dependence. Ranking/cost analyses are descriptive. Fitting and threshold selection use training/validation, while successive exploratory analyses share the fixed test partition.

5 Results

5.1 RQ1: Coverage and Output Semantics

Only 2,220 of 14,637 test records (15.17%) had readable CXR files, compared with 2,232 eligible references. Mortality and LOS prevalence were higher in this subset than among records without

Table 1: Clean results on 14,637 test records. S uses five expert logits; U adds engineered features; R adds perturbation training to U. All variants use class-balanced fitting and raw sigmoid scores. Lower Brier and 15-bin ECE are better.

Outcome	Model	AUROC	AUPRC	Brier	ECE	Bal. acc.
Mortality	Clinical stack	0.8676	0.5360	0.1459	0.2224	0.7844
Mortality	Scores-only S	0.8669	0.5328	0.1398	0.2224	0.7844
Mortality	CARE-Fuse-U	0.8674	0.5344	0.1405	0.2164	0.7842
Mortality	CARE-Fuse-R	0.8675	0.5350	0.1399	0.2145	0.7840
LOS	Clinical stack	0.8305	0.6869	0.1689	0.1113	0.7455
LOS	Scores-only S	0.8300	0.6878	0.1686	0.1128	0.7426
LOS	CARE-Fuse-U	0.8306	0.6876	0.1685	0.1124	0.7456
LOS	CARE-Fuse-R	0.8301	0.6867	0.1691	0.1103	0.7446

Table 2: Five-draw mean AUROC at 50% stress for all four families. Differences use unrounded means. Final rows give the minimum condition mean for each model across four families and three nonzero rates; the minimizing condition can differ by model. R-minus-S comparisons are descriptive.

Outcome	Stress / summary	Clinical	S	U	R	R minus S
Mortality	GRU removal	0.8552	0.8627	0.8536	0.8638	+0.0012
Mortality	Report removal	0.8665	0.8659	0.8663	0.8664	+0.0005
Mortality	Temporal/report removal	0.8570	0.8615	0.8549	0.8629	+0.0014
Mortality	Auxiliary conflict	0.8676	0.8402	0.8400	0.8634	+0.0233
LOS	GRU removal	0.7088	0.8255	0.7146	0.8262	+0.0007
LOS	Report removal	0.8293	0.8292	0.8294	0.8294	+0.0002
LOS	Temporal/report removal	0.7189	0.8232	0.7240	0.8253	+0.0021
LOS	Auxiliary conflict	0.8305	0.8293	0.8299	0.8299	+0.0006
Mortality	Minimum condition mean	0.8552	0.8402	0.8400	0.8629	—
LOS	Minimum condition mean	0.7088	0.8232	0.7146	0.8253	—

readable images. Image replacement reduced LLaVA-Med mortality AUROC by 0.0209 (95% CI [0.0002, 0.0410]) and AUPRC by 0.0226 ([0.0070, 0.0399]), but changed only 0.14% of hard answers. Thus image sensitivity in ranking need not produce a different thresholded decision. Luna and Sol had complete parse coverage but only partial candidate-score coverage. Appendix E supplies the full coverage table, output metrics, and reconstruction comparisons.

5.2 RQ2: Comparator Choice Changes the Robustness Conclusion

5.2.1 Clean ranking is similar across variants; calibration error persists.

Clean discrimination lies in a narrow range across the clinical stack, scores-only S, CARE-Fuse-U, and CARE-Fuse-R (Table 1). Mortality AUROC ranges from 0.8669 to 0.8676 and LOS AUROC from 0.8300 to 0.8306. CARE-Fuse-R achieved AUROC/AUPRC 0.8675/0.5350 for mortality and 0.8301/0.6867 for LOS. For mortality, CARE-Fuse-R had Brier score 0.1399 and ECE 0.2145; all four Brier scores exceeded the observed-prevalence reference of approximately 0.1022. LOS CARE-Fuse-R ECE was 0.1103. These class-balanced score pipelines provide discrimination alongside substantial calibration error on the original event distribution.

5.2.2 Simple fusion narrows the apparent benefit under output loss.

With joint temporal/report-output removal applied to 50% of records, mean LOS AUROC was 0.8253 for R, 0.8232 for S, and 0.7240 for U (Table 2). The R-minus-U difference of 0.1013 becomes 0.0021 relative to S. Mortality follows the same ordering, with R-minus-S 0.0014. Thus much of the apparent gain over U concerns the fragility of that engineered comparator under this removal rule. The archived first-draw paired bootstraps quantify the R-minus-U removal contrasts: 0.0075 (95% CI [0.0049, 0.0102]) for mortality and 0.1023 ([0.0949, 0.1092]) for LOS. These intervals

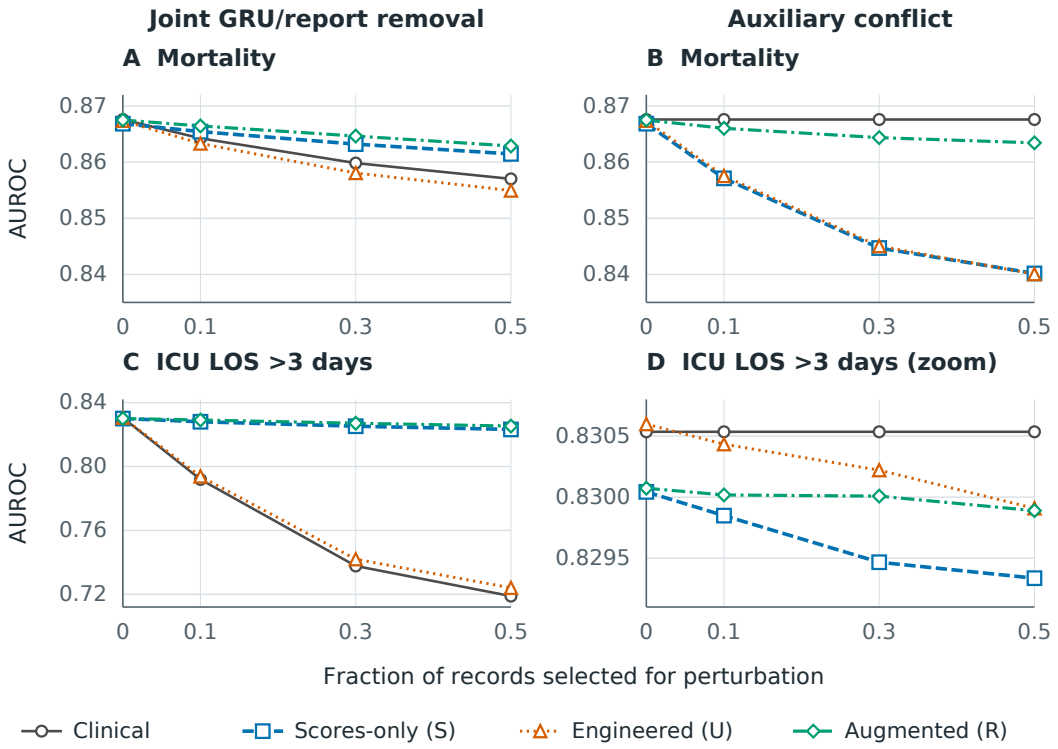


Figure 2: Comparator response across joint removal and auxiliary conflict. Nonzero rates show five-draw means; zero is the single clean result. Vertical scales differ, with a labeled zoom for LOS conflict. These are descriptive curves; draw variation and all four stress families are included in the supplement.

concern the engineered comparator within one perturbation draw; the smaller S contrasts have no corresponding paired record intervals in the archive.

5.2.3 The broader stress comparison exposes different vulnerabilities.

Under 50% auxiliary conflict, mortality mean AUROC was 0.8402 for S and 0.8400 for U, versus 0.8634 for R. The clinical stack, which excludes the altered score, remained at 0.8676. Its first-draw advantage over R was 0.0043 (R-minus-clinical 95% CI [-0.0061, -0.0023]). LOS conflict differences were much smaller (Figure 2).

Across all 12 tested conditions, R had the highest minimum condition-mean AUROC for each outcome. S's mortality minimum occurred under auxiliary conflict, whereas the clinical stack's occurred under GRU removal. R's minima occurred under joint removal. R therefore attains the highest worst-condition mean AUROC within the tested set, while the simple baselines identify individual conditions with similar observed performance at lower complexity.

Maximum gaps to the condition-specific best were 0.0086, 0.0274, 0.0276, and 0.0042 for mortality and 0.1174, 0.0021, 0.1117, and 0.0006 for LOS, in clinical/S/U/R order. R had the smallest maximum gap for both outcomes. Its largest gaps occurred under 50% auxiliary conflict, whereas its lowest absolute AUROC occurred under joint removal. The two summaries identify different limiting conditions while supporting R's relative consistency across the tested set.

Table 3: Selected errors at a 10% budget using CARE-Fuse-U ($m = 1{,}464$). FP and FN are counts; capture percentages divide by all errors or all false negatives for the outcome.

Outcome	Policy	Selected FP	Selected FN	Error capture	FN capture
Mortality	Threshold proximity	618	68	22.52%	17.80%
Mortality	Disagreement	677	8	22.49%	2.09%
Mortality	Combined	722	51	25.38%	13.35%
LOS	Threshold proximity	482	215	17.29%	23.24%
LOS	Disagreement	351	12	9.01%	1.30%
LOS	Combined	614	131	18.48%	14.16%

5.3 RQ3: Error Objectives Change the Preferred Ranking

5.3.1 Higher total-error capture favors a different error composition.

The CARE-Fuse-U mortality predictions contained 2,664 false positives and 382 false negatives; LOS predictions contained 3,106 and 925, respectively. False positives comprised 87.46% and 77.05% of total errors. At a 10% budget, each policy selected 1,464 records (Table 3). Combined captured 25.38% of mortality errors and 18.48% of LOS errors, corresponding to 2.54 and 1.85 times error enrichment. Random selection would capture approximately 10% in expectation.

Relative to threshold proximity, combined selected 104 additional mortality false positives and 17 fewer false negatives; for LOS, it selected 132 additional false positives and 84 fewer false negatives. These exchanges increased total-error capture by 2.86 and 1.19 percentage points while reducing false-negative capture. At 30% mortality selection, disagreement had the largest total-error capture: 67.10%, compared with 65.53% for combined.

5.3.2 Cost ratios locate the policy crossovers.

At 10% mortality selection, combined's selected weighted-error total is $722 + 51\lambda$, and threshold proximity's is $618 + 68\lambda$. Equality occurs at $\lambda = 104/17 \approx 6.118$: combined has lower retained weighted error below this ratio, and proximity has lower error above it. LOS crosses at $\lambda = 11/7 \approx 1.571$, using selected counts 614 FP/131 FN for combined and 482 FP/215 FN for proximity (Figure 3). At this budget, disagreement selects fewer errors of both types than combined.

The cost boundary also depends on the budget (Table 4). With equal FP/FN costs, combined has the lowest retained weighted error in seven of eight outcome/budget settings; disagreement leads for mortality at 30%. At sufficiently high relative FN cost, proximity leads in all eight. For example, increasing λ from one to two in the 10% LOS setting changes proximity-versus-combined retained weighted-error totals from 3,334 versus 3,286 to 4,044 versus 4,080, reversing the ordering.

The gate's threshold-selection metric supplies a useful reference objective. Maximizing validation balanced accuracy is equivalent to minimizing $\mathrm{FP} + \lambda_{\mathrm{BA}}\mathrm{FN}$ with $\lambda_{\mathrm{BA}} = N_{\mathrm{val},-}/N_{\mathrm{val},+}$. Validation counts yield approximately 7.66 for mortality and 2.17 for LOS. Both exceed the 10% policy crossovers, favoring proximity over combined when carrying those fixed weights into retained-error evaluation. Thus maximizing total-error capture changes the error objective used to select the classifier threshold. The validation-derived weights remain fixed across policies. This weighted retained error differs from retained-set balanced error, whose class denominators change with selection (Narasimhan et al., 2024).

At the validation-derived weights, proximity lowers weighted error relative to combined by 1.991 per 1,000 retained records for mortality (0.586% relative to combined) and 3.786 for LOS (1.184%) at 10% selection. At 30%, the corresponding reductions are 15.807 and 11.323 per 1,000; at 5% LOS, combined instead has lower weighted error by 0.662 per 1,000. The supplement reports all eight outcome/budget differences in units of FP cost, normalized to one.

5.3.3 Selection changes accuracy and sensitivity through case mix.

At 10% mortality selection, disagreement removes 665 true positives and eight false negatives. Retained event prevalence falls from 11.56% to 7.74%. Ordinary retained error decreases from

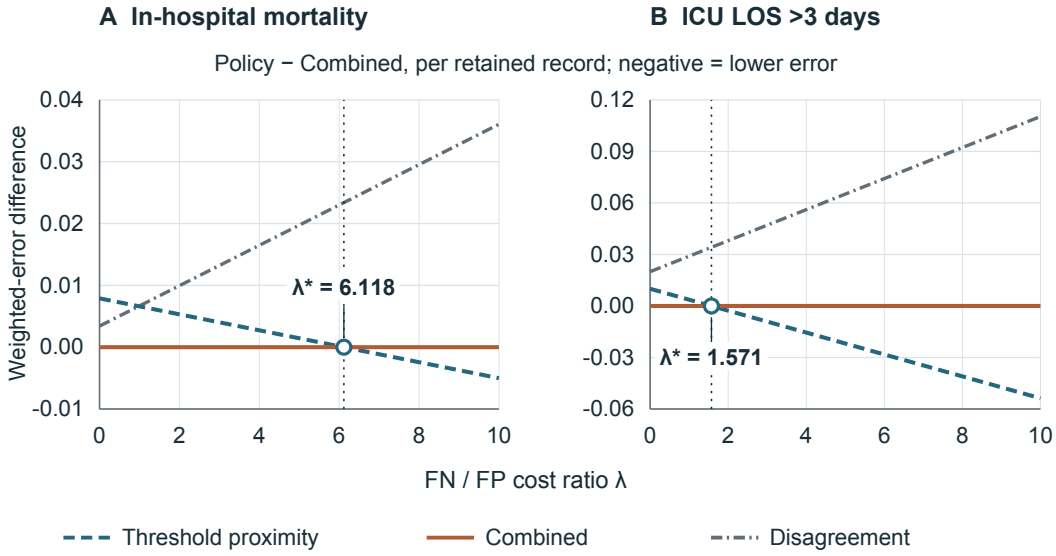


Figure 3: Cost sensitivity at 10% selection using CARE-Fuse-U, with 13,173 retained records. Curves show retained weighted error per record minus Combined: negative values favor the plotted policy. Crossovers are calculated from the aggregate confusion counts.

Table 4: Lowest retained weighted error among the three evaluated rankings at each fixed budget. P: threshold proximity; C: combined; D: disagreement. Adjacent policies tie at each exact boundary.

Outcome	Budget	Minimum retained weighted error for $\lambda > 0$
Mortality	5%	C for $\lambda < 46/13$ (3.538); P above
Mortality	10%	C for $\lambda < 104/17$ (6.118); P above
Mortality	20%	C for $\lambda < 27/5$ (5.400); P above
Mortality	30%	D for $\lambda < 55/31$ (1.774); C to $290/59$ (4.915); P above
LOS	5%	C for $\lambda < 59/23$ (2.565); P above
LOS	10%	C for $\lambda < 11/7$ (1.571); P above
LOS	20%	C for $\lambda < 231/148$ (1.561); P above
LOS	30%	C for $\lambda < 178/109$ (1.633); P above

20.81% to 17.92%, while conditional false-negative rate rises from 22.58% to 36.70%. The fixed predictions become a more accurate but less sensitive retained set through selection alone.

For LOS, disagreement at the same budget increases both ordinary retained error (27.54% to 27.84%) and conditional false-negative rate (19.44% to 24.46%). Its total-error capture, 9.01%, is below the approximately 10% selected fraction. Retained confusion counts and outcome prevalence explain these movements alongside aggregate error capture.

6 Discussion and Limitations

The comparison must match the operation. Coverage defines the evaluated population; continuous ranking can change while hard answers remain fixed; raw class-balanced scores can discriminate without supporting probability interpretation. The observed calibration error calls for matched calibration comparisons rather than attributing it uniquely to weighting or fusion (van den Goorbergh et al., 2022). Simple S fusion substantially narrows the apparent robustness gain over U under joint output removal, while the clinical comparator exposes sensitivity to auxiliary conflict. R is relatively consistent across the tested conditions without winning each condition. Training and testing share failure families, so the result does not establish robustness to arbitrary shifts.

Error allocation further depends on the objective. Combined captures more FP, whereas proximity captures more FN. Carrying validation balanced-accuracy weights into the retained-cost comparison reverses the equal-error preference at 10% selection. Retained accuracy can also rise while sensitivity falls through case selection alone. A selection component therefore needs selected error composition, retained confusion counts, and an explicit error objective. The analysis does not measure clinician performance or the benefit of review. Appendix G expands these implications.

The retrospective single-center archive has stay-level partitions, possible repeated-admission overlap, and no common 24-hour landmark check. Missing patient mappings and availability timestamps prevent retrospective correction. New synthetic checks validate metadata declarations and hashes, not actual historical payloads or upstream encoder-training provenance. Stress modifies expert scores and flags, not raw modalities; fitting seeds, regularization, and augmented training size also differ. Comparisons describe fitted variants, not an isolated causal mechanism. Ranking uses U and in-sample validation reference distributions. Marginal aggregates omit selection overlap, tie sensitivity, paired ranking intervals, and subgroup contrasts. Deployment requires independent costs and validation; comparison across budgets additionally requires a selection cost.

7 Conclusion

CARE-Fuse shows how output definitions, comparators, and error objectives change multimodal prediction evaluation. Threshold-selection weights favor a different 10% selection policy from equal-error capture in this archive. Simple fusion narrows the apparent robustness gain under output loss, while perturbation-trained fusion has the highest minimum condition mean across the tested stresses. Score diagnostics and error decomposition make these component-evaluation choices explicit.

AI USE STATEMENT

AI tools were used to assist with polishing the manuscript, including improving grammar, clarity, sentence structure, paragraph organization, and consistency of academic terminology. They also supported refinement of English phrasing to improve readability and presentation.

ETHICS STATEMENT

The source-resource documentation reports institutional review of data collection and resource creation, waived informed consent, and approved sharing. This aggregate audit involves no participant recruitment or prospective intervention; IRB review was not required for this study. Access and reuse remain governed by PhysioNet conditions (Pollard et al., 2026). The shared materials contain aggregates, configurations, and code rather than individual patient records or outputs. Archived API settings include `store=false`, but do not establish account-level data-processing conditions. The findings do not establish clinical benefit or outcomes of human review.

REPRODUCIBILITY STATEMENT

The companion archive contains aggregate tables, fitting configurations, protocols, analysis code, claim mappings, and hashes. Its verification command checks frozen inputs, reruns aggregate analyses, compares generated tables and summaries, and exercises failure cases. Appendix F details the pipelines; Appendix C records configurations and seeds. Historical model reruns require authorized source-data access and the recorded environment. The cohort tools and synthetic tests support subsequent runs and do not certify the unavailable historical inputs.

REFERENCES

Baraa Al Jorf and Farah E. Shamout. MedPatch: Confidence-guided multi-stage fusion for multimodal clinical data. In *Proceedings of the 10th Machine Learning for Healthcare Conference*, volume 298 of *Proceedings of Machine Learning Research*, pp. 1–29. PMLR, 2025. URL https://proceedings.mlr.press/v298/jorf25a.html.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. Association for Computing Machinery, 2019. doi: 10.1145/3290605.3300233.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16. Association for Computing Machinery, 2021. doi: 10.1145/3411764.3445717.

Mélanie Bernhardt, Fabio De Sousa Ribeiro, and Ben Glocker. Failure detection in medical image classification: A reality check and benchmarking testbed. *Transactions on Machine Learning Research*, October 2022. URL https://openreview.net/forum?id=VBHuLfnOMf.

Ángel Alexander Cabrera, Erica Fu, Donald Bertucci, Kenneth Holstein, Ameet Talwalkar, Jason I. Hong, and Adam Perer. Zeno: An interactive framework for behavioral evaluation of machine learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3544548.3581268. URL https://doi.org/10.1145/3544548.3581268.

Luís Felipe Cattelan and Danilo Silva. How to fix a broken confidence estimator: Evaluating post-hoc methods for selective classification with deep neural networks. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 547–584. PMLR, 2024. URL https://proceedings.mlr.press/v244/cattelan24a.html.

Chris Drummond and Robert C. Holte. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006. doi: 10.1007/s10994-006-8199-5. URL https://link.springer.com/article/10.1007/s10994-006-8199-5.

Shaza Elsharief, Saeed Shurrab, Baraa Al Jorf, Leopoldo Julian Lechuga Lopez, Krzysztof J. Geras, and Farah E. Shamout. MedMod: Multimodal benchmark for medical prediction tasks with electronic health records and chest x-ray scans. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pp. 781–803. PMLR, 2025. URL https://proceedings.mlr.press/v287/elsharief25a.html.

Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023. URL https://jmlr.org/papers/v24/21-0048.html.

Yanjun Gao, Skatje Myers, Shan Chen, Dmitriy Dligach, Timothy Miller, Danielle S. Bitterman, Guanhua Chen, Anoop Mayampurath, Matthew M. Churpek, and Majid Afshar. Uncertainty estimation in diagnosis generation from large language models: next-word probability is not pre-test probability. *JAMIA Open*, 8(1):ooae154, 2025. doi: 10.1093/jamiaopen/ooae154.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. doi: 10.1145/3458723. URL https://dl.acm.org/doi/10.1145/3458723.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152. PMLR, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Akshat Gupta and Jianguo Liu. Selective prediction and the persistence illusion: A diagnostic decomposition of VIX regime classification. *Journal of Risk and Financial Management*, 19(8), August 2026. doi: 10.3390/jrfm19080609. URL https://www.mdpi.com/1911-8074/19/8/609.

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. MedFuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Proceedings of Machine Learning for Healthcare*, volume 182 of *Proceedings of Machine Learning Research*, pp. 479–503. PMLR, 2022. URL https://proceedings.mlr.press/v182/hayat22a.html.

Nikkie Hooman, Zhongjie Wu, Eric C. Larson, and Mehak Gupta. Multimodal routing for interpretable, robust, and auditable clinical prediction, 2026. URL https://arxiv.org/abs/2607.09982v1. Author manuscript, version 1, 10 July 2026.

Sarah Jabbour, David Fouhey, Nikola Banovic, Stephanie D. Shepard, Ella Kazerooni, Michael W. Sjoding, and Jenna Wiens. Selective prediction reduces the negative effects of automation bias overall but increases false negatives, 2026. URL https://arxiv.org/abs/2508.07617v2. Preprint, version 2, revised 11 August 2026; first posted 11 August 2025.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. PhysioNet, January 2023a. URL https://physionet.org/content/mimiciv/2.2/. Version 2.2.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes. PhysioNet, January 2023b. URL https://physionet.org/content/mimic-iv-note/2.2/. Version 2.2.

Alistair Johnson, Matthew Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR-JPG - chest radiographs with structured labels. PhysioNet, March 2024. URL https://physionet.org/content/mimic-cxr-jpg/2.1.0/. Version 2.1.0.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. doi: 10.1038/s41597-019-0322-0.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023c. doi: 10.1038/s41597-022-01899-x.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. Selective classification can magnify disparities across groups. In *International Conference on Learning Representations*, pp. 1–34. International Conference on Learning Representations, 2021. URL https://arxiv.org/abs/2010.14134.

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

Ranit Karmakar and Jayita Chatterjee. What single-prompt accuracy misses: A multi-variant reliability audit of language models, 2026. URL https://arxiv.org/abs/2605.02038. Preprint.

Attila Kovari. Stable discrimination can hide reliability failures in AI decision support under distribution shift and changing target definitions. *Computers*, 15(9), 2026. doi: 10.3390/computer s15090560. URL https://www.mdpi.com/2073-431X/15/9/560.

Kwanhyung Lee, Soojeong Lee, Sangchul Hahn, Heejung Hyun, Edward Choi, Byungeun Ahn, and Joohyung Lee. Learning missing modal electronic health records with unified multi-modal data embedding and modality-aware attention. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pp. 423–442. PMLR, 2023. URL https://proceedings.mlr.press/v219/lee23a.html.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28541–28564. Curran Associates, Inc., 2023. doi: 10.52202/075280-1240. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306. IEEE, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. ECBD: Evidence-centered benchmark design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16349–16365, Bangkok, Thailand, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.861. URL https://aclanthology.org/2024.acl-long.861/.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022. IEEE, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html.

L Julian Lechuga Lopez, Farah E Shamout, and Tim G J Rudner. An empirical analysis of calibration and selective prediction in multimodal clinical condition classification. In *Proceedings of the 7th Conference on Health, Inference, and Learning*, volume 333 of *Proceedings of Machine Learning Research*, pp. 794–833. PMLR, 2026. URL https://proceedings.mlr.press/v333/lopez26a.html.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229. Association for Computing Machinery, 2019. doi: 10.1145/3287560.3287596. URL https://dl.acm.org/doi/10.1145/3287560.3287596.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, Neha Gupta, and Sanjiv Kumar. Learning to reject meets long-tail learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c4f129179494c1ea14b63fc0019f3095-Abstract-Conference.html.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, August 2026. doi: 10.1038/s44360-026-00096-z. URL https://doi.org/10.1038/s44360-026-00096-z.

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–52. Association for Computing Machinery, 2021. doi: 10.1145/3411764.3445315. URL https://dl.acm.org/doi/10.1145/3411764.3445315.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44. Association for Computing Machinery, 2020. doi: 10.1145/3351095.3372873. URL https://dl.acm.org/doi/10.1145/3351095.3372873.

Fatih Sağlam, Ünsal Özgen, Alper Uygun, Oğuz Salih Dinçer, and Canan Albayrak. Selective classification under imbalance in multiclass settings: A novel metric for bias-aware risk–coverage evaluation. *Journal of Biomedical Informatics*, 181, 2026. doi: 10.1016/j.jbi.2026.105084.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. "Everyone Wants to Do the Model Work, Not the Data Work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. Association for Computing Machinery, 2021. doi: 10.1145/3411764.3445518. URL https://dl.acm.org/doi/10.1145/3411764.3445518.

Annika Schneider, Tommy Rochussen, Joshua Stiller, and Vincent Fortuin. Decision-aligned evaluation of uncertainty quantification, 2026. URL https://arxiv.org/abs/2606.26990v1. Preprint, version 1, 25 June 2026.

Venkatesh Sivaraman, Anika Vaishampayan, Xiaotong Li, Brian R. Buck, Ziyong Ma, Richard D. Boyce, and Adam Perer. Tempo: Helping data scientists and domain experts collaboratively specify predictive modeling tasks. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3706598.3713664. URL https://doi.org/10.1145/3706598.3713664.

Akshay Swaminathan, Ivan Lopez, William Wang, Ujwal Srivastava, Edward Tran, Aarohi Bhargava-Shah, Janet Y. Wu, Alexander L. Ren, Kaitlin Caoili, Brandon Bui, Layth Alkhani, Susan Lee, Nathan Mohit, Noel Seo, Nicholas Macedo, Winson Cheng, Charles Liu, Reena Thomas, Jonathan H. Chen, and Olivier Gevaert. Selective prediction for extracting unstructured clinical data. *Journal of the American Medical Informatics Association*, 31(1):188–197, 2024. doi: 10.1093/jamia/ocad182. URL https://academic.oup.com/jamia/article/31/1/188/7285661.

Vinitra Swamy, Malika Satayeva, Jibril Frej, Thierry Bossy, Thijs Vogels, Martin Jaggi, Tanja Käser, and Mary-Anne Hartley. MultiMoDN—multimodal, multi-task, interpretable modular networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28115–28138. Curran Associates, Inc., 2023. doi: 10.52202/075280-1221. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5951641ad71b0052cf776f9b71f18932-Abstract-Conference.html.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems*, volume 37, pp. 2323–2347. Curran Associates, Inc., 2024. doi: 10.52202/079017-0076. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/047c84ec50bd8ea29349b996fc64af4b-Abstract-Conference.html.

Ruben van den Goorbergh, Maarten van Smeden, Dirk Timmerman, and Ben Van Calster. The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, 29(9):1525–1534, 2022. doi: 10.1093/jamia/ocac093.

Andrew Wang, Jiashuo Zhang, and Michael Oberst. Revisiting performance claims for chest x-ray models using clinical context. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 333 of *Proceedings of Machine Learning Research*, pp. 493–536. PMLR, 2026. URL https://proceedings.mlr.press/v333/wang26c.html.

Chenyang Yang, Tesi Xiao, Michael Shavlovsky, Christian Kästner, and Tongshuang Wu. Orbit: A framework for designing and evaluating multi-objective rankers. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 1093–1106, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3708359.3712092. URL https://doi.org/10.1145/3708359.3712092.

Wenfang Yao, Kejing Yin, William K. Cheung, Jia Liu, and Jing Qin. DrFuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16416–16424. AAAI Press, 2024. doi: 10.1609/aaai.v38i15.29578. URL https://ojs.aaai.org/index.php/AAAI/article/view/29578.

Kejing Yin, Haizhou Xu, Wenfang Yao, Chen Liu, Zijie Chen, Yui Haang Cheung, William K. Cheung, and Jing Qin. When does multimodal learning help in healthcare? a benchmark on EHR and chest x-ray fusion, 2026. Preprint.

Kunyu Yu, Rui Yang, Jingchi Liao, Siqi Li, Huitao Li, Irene Li, Yifan Peng, Rishikesan Kamaleswaran, and Nan Liu. Benchmarking foundation models with multimodal public electronic health records. *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2025. doi: 10.1109/JBHI.2025.3645076. URL https://doi.org/10.1109/JBHI.2025.3645076. Early access.

A Initial Fusion Experiment

The initial experiment used the same 14,637-record test population but a gate-fitting configuration preceding the matched mechanism refits. It achieved mortality AUROC/AUPRC 0.8674/0.5344 and LOS 0.8310/0.6878. This initial configuration supplies the single-expert and ablation comparisons below. Five-seed AUROC ranges were 0.8665–0.8674 for mortality and 0.8299–0.8310 for LOS, describing initialization and fitting variability on the fixed split.

The initial fusion was compared with the laboratory expert for mortality and the GRU expert for LOS, which had the highest observed test AUROC among the individual experts. Paired 5,000-draw intervals condition on these fixed comparator identities. Gains were +0.0428 mortality AUROC (95% CI [0.0368, 0.0491]), +0.0671 mortality AUPRC ([0.0573, 0.0770]), +0.0386 LOS AUROC ([0.0339, 0.0433]), and +0.0960 LOS AUPRC ([0.0860, 0.1070]). The LOS AUPRC values were 0.6237 for vital signs and 0.5918 for GRU; the AUPRC contrasts use the same fixed comparators.

Ablations identified structured features as the backbone. Removing laboratories reduced AUROC by 0.0416 for mortality and 0.0164 for LOS; removing the vital-sign/demographic/admission branch reduced it by 0.0139 and 0.0191. Report and auxiliary branches supplied smaller task-dependent changes. These ablations identify structured input groups as the largest contributors to discrimination in the initial fusion configuration.

B Additional Reconstruction and Output Diagnostics

The encoder reconstruction combines a 1,024-dimensional GRU representation of 24 hourly physiological bins, a 768-dimensional RadBERT representation, and a 1,024-dimensional Swin image representation using logistic outcome heads. The complete-case cohort contains 11,442 records: 8,076 training, 1,119 validation, and 2,247 test. Three initialization seeds hold the cohort and partition fixed. A separate coverage-aware experiment uses a structured-only model for the full test partition and substitutes a complete-feature multimodal model on its eligible subset. Binning, representations, missingness handling, and cohort selection differ from the source benchmark. Supporting temporal-foundation-model experiments use MOMENT (Goswami et al., 2024).

The Qwen diagnostics provide another example of why ranking and decision channels must remain distinct. In the validation-calibration protocol, raw-score test AUROC was 0.6113 for mortality and 0.5395 for LOS despite almost entirely negative hard decisions. Calibration fitted on the 7,318-record validation partition reduced test mortality ECE from 0.0433 to 0.0170. LOS validation AUROC was approximately random (0.4941), producing an almost constant monotone calibrator. Candidate-order changes affected rankings, so these findings remain specific to the frozen prompt and scoring protocol.

Sol also had an earlier 1,600-record validation diagnostic. Luna and Sol had 100% parse coverage in the completed validation and test runs. Their candidate-score coverage across split and outcome ranged from 0.007% to 9.03% for Luna and 0.15% to 26.77% for Sol. On the same test cohort, Sol's LOS hard-decision balanced accuracy exceeded Luna's by 0.0121, sensitivity by 0.0492, and F1 by 0.0633, while specificity decreased by 0.0249. Mortality differences were small because both produced almost entirely negative hard answers. These values describe the hard-decision operating points on the same test cohort.

C Implementation Details and Aggregate Evidence

The source paper describes an 80/20 train/test split (Yu et al., 2025); its released classification code reserves 20% of metadata rows for testing, then 12.5% of the remainder for validation. The archived cohort builder likewise assigns stored split labels before sorting rows by stay identifier. The main fusion reader subsequently reindexes the sorted rows and reapplies the random-state-0 rule, whereas the encoder reconstruction consumes the stored labels. These are stay-level procedures. Equal seeds and proportions do not establish equal membership after row reordering; the archived QA checks establish agreement with the main fusion reader's record order and labels.

Structured vital extraction uses elapsed chart time in $[0, 24)$ hours and caps events at ICU exit. The author-compatible temporal builder applies $[0, 24)$ without an explicit exit cap. Laboratory extraction includes chart time in $[0, 24]$ without an explicit exit cap, while CXR acquisition uses

$[0, \min(24, \text{ICU duration})]$. The archived CXR parser substitutes midnight for a missing clock value. Linked report text is selected by study time without an independent report-availability check. Thus an acquisition-window condition does not establish when the report content became available.

The `cohort_validation/` tools join stays and admissions by identifier, set prediction to ICU admission plus exactly 24 hours, and exclude ICU exit, hospital discharge, or in-hospital death at or before prediction. SHA-256 assignment of the seed and patient key gives an order-independent patient partition with approximate 70/10/20 proportions. Preparation produces a separate cohort and a `prepared_not_validated` report. The lineage checker requires each declared modality's event range and latest actual content availability to fall within the observation window; derived inputs inherit the latest availability of all content they used. Missing lineage remains incomplete. Current fitting entrypoints require a passing metadata contract, agreement with the actual row partitions and any stored split labels, and hashes for the expected input files before loading model libraries. These checks bind the declared lineage to file bytes; they do not inspect whether the payloads agree with declared modality presence. Reports contain aggregate diagnostics and input hashes; identifier-bearing rows remain local. Synthetic tests exercise the contract; auditing original experiments requires their complete metadata and provenance.

The gates use clipping constant $\epsilon = 10^{-5}$ and population standard deviations ($\text{ddof} = 0$). Main four-model ECE uses 15 equal-width bins; initial-fusion calibration, VLM beyond-binary diagnostics, Qwen monotone calibration, and the separate laboratory extension use ten. The reliability gate selected $C = 0.001$ and threshold 0.481417635 for mortality, and $C = 0.01$ and threshold 0.482651041 for LOS. Augmentation seeds are 20260921–20260923 with scenario offsets of $1000j$; test seeds are 20260911–20260915 with offsets $1000t + 100j$, where t and j index the outcome and scenario in the archived implementation. Triage CDF ranks use right-inclusive comparisons, with equal weights across the three percentile components.

The supplement's `claim_evidence_map.csv` maps claims to aggregate rows. Its analysis directories provide cost calculations, comparator summaries, and 1,792-cell agreement checks. The `protocol/` directory preserves API settings and eight synthetic extraction cases. The README separates aggregate reproduction from new data runs; manifests record hashes.

D Extended Related Work

D.1 Fusion with Incomplete and Conflicting Evidence

The source benchmark evaluates modality encoders and multimodal learners on ICU outcomes and already reports limited CXR availability and greater modality completeness among adverse-outcome groups (Yu et al., 2025). Its representations include temporal encoders, text models, and chest-image encoders such as Swin (Liu et al., 2021). CARE-Fuse examines additional properties of the archived outputs: the gap between referenced and locally readable images, and the difference between continuous ranking changes and hard-decision changes under image replacement. Coverage imbalance supplies the context for these output-level comparisons.

Methods for incomplete clinical evidence address several levels of the modeling pipeline. MedFuse combines paired and partially paired EHR–CXR data (Hayat et al., 2022); modality-aware embedding and attention represent irregular and incomplete EHR inputs (Lee et al., 2023); MultiModN supports modular prediction under changing missingness (Swamy et al., 2023); and DrFuse separates shared and modality-specific representations to address missingness and inconsistency (Yao et al., 2024). MedPatch uses confidence-guided fusion for incomplete clinical data (Al Jorf & Shamout, 2025). MedMod standardizes five EHR–CXR prediction tasks (Elsharief et al., 2025), while contextual evaluation shows how chest-image performance changes across populations defined by prior clinical information (Wang et al., 2026).

The 2026 CareBench preprint compares simple and advanced EHR–CXR fusion under complete and missing inputs, including modality imbalance and subgroup performance (Yin et al., 2026). Its patient-level partitions and 48-hour mortality and LOS specifications differ from the source benchmark's 24-hour input definition. Hooman et al. audit learned multimodal routes by disabling modality-dependent routes at inference (Hooman et al., 2026). Their mortality experiment uses paired trimodal records from the first 48 hours. CARE-Fuse focuses on scalar expert outputs within the

frozen archive: the five-score comparator measures response to the specified output replacements using expert scores alone, while U and R additionally use engineered availability and disagreement features. This comparison complements representation-learning approaches by fixing the score interface and making the comparator's input features explicit.

D.2 Score Interpretation and Selective Prediction

Calibration measures agreement between scores and outcome frequency, whereas AUROC measures ranking (Guo et al., 2017). Candidate-normalized token scores and sigmoid outputs require empirical calibration assessment before they can support probability interpretation. In three EHR diagnosis tasks, Gao et al. found that token-logit and verbalized estimates underperformed a classifier trained on structured data and showed poor calibration (Gao et al., 2025). A 2026 preprint reports that label-set normalization and answer parsing alter calibration and accuracy conclusions across language-model benchmarks (Karmakar & Chatterjee, 2026). These results place the output rule within the evaluation specification.

Selective classification formalizes abstention from some inputs (Geifman & El-Yaniv, 2017); learning to defer additionally models when a downstream expert should decide (Mozannar & Sontag, 2020). Franc et al. characterize optimal rejection strategies under cost, risk, and coverage formulations (Franc et al., 2023). In clinical note abstraction, Swaminathan et al. optimize asymmetric rejection regions using false-positive, false-negative, and abstention costs (Swaminathan et al., 2024).

Evaluation depends on the chosen objective as well as the selection rule. Failure-detection benchmarks distinguish prediction-correctness discrimination from risk–coverage evaluation (Bernhardt et al., 2022), and Cattelan and Silva show how confidence estimation changes selection performance for fixed classifiers (Cattelan & Silva, 2024). Traub et al. distinguish conditional selective risk from generalized risk and demonstrate that changing a multi-threshold metric can change method rankings (Traub et al., 2024). Narasimhan et al. establish that rejectors suited to ordinary error can be suboptimal for balanced error, whose class weights depend on the rejection rule (Narasimhan et al., 2024). A 2026 preprint by Schneider et al. formalizes alignment between uncertainty metrics and downstream utility families and proposes prior-weighted utility metrics (Schneider et al., 2026).

Empirical comparisons expose the consequences of these choices. Gupta and Liu report that varying FP/FN costs reverses the preference between random-forest selective predictions and a persistence comparator in VIX regime classification (Gupta & Liu, 2026). Jones et al. show that selective classification can increase group disparities despite improving average accuracy (Jones et al., 2021). In imbalanced multiclass tasks, class-aware evaluation exposes differences in rejection across outcome classes that aggregate risk–coverage metrics can conceal (Sağlam et al., 2026). Lechuga López et al. document class-dependent calibration and selective-performance failures in paired EHR–CXR condition classification (Lopez et al., 2026).

Human responses introduce a further evaluation question. A recent preprint by Jabbour et al. reports that withholding inaccurate AI predictions partly restores clinicians' treatment accuracy while increasing false negatives relative to a no-AI baseline (Jabbour et al., 2026).

CARE-Fuse provides a fixed-classifier empirical analysis of these evaluation choices. For two binary outcomes, it compares three archived rankings through selected and retained FP/FN counts, derives their fixed-budget cost crossovers, and locates those crossovers relative to weights implied by validation threshold selection. The measured outcome is error allocation with the predictions unchanged.

HCI evaluation systems make modeling decisions inspectable: Zeno supports behavioral evaluation (Cabrera et al., 2023), Tempo makes temporal prediction specifications explicit (Sivaraman et al., 2025), and Orbit organizes ranker design around objectives and their tradeoffs (Yang et al., 2025). CARE-Fuse supplies computational checks for the prediction components such systems may expose: input availability, score semantics, comparator choice, and error objectives. These checks specify component behavior that can inform subsequent studies of joint human–AI outcomes (Bansal et al., 2021; Poursabzi-Sangdeh et al., 2021).

Table 5: Test-set evidence availability. Mortality and LOS columns give observed outcome prevalence.

Evidence state	Records	Test share	Mortality	LOS >3d
All test records	14,637	100.00%	0.1156	0.3251
CXR reference present	2,232	15.25%	0.1514	0.4368
Readable image	2,220	15.17%	0.1518	0.4365
No readable image	12,417	84.83%	0.1091	0.3052

The reference and readable-image rows overlap. Twelve referenced files were unavailable locally; 12,405 records had no eligible reference.

D.3 Context and Reproducibility in Model Evaluation

MIMIC events, reports, and images are incompletely paired (Johnson et al., 2023c; 2019), so complete-case filtering changes the evaluated population. Datasheets, model cards, end-to-end auditing, and research on data cascades motivate documenting these dependencies across a modeling pipeline (Gebru et al., 2021; Mitchell et al., 2019; Raji et al., 2020; Sambasivan et al., 2021).

Evidence-centered benchmark design distinguishes observed model responses from the capabilities inferred from them (Liu et al., 2024b). Kapoor and Narayanan distinguish repeatability from valid prediction claims, including sampling and preprocessing leakage (Kapoor & Narayanan, 2023). CRIT-AID jointly audits probability mappings, selective operating rules, and conformal outputs under source-to-target transport (Kovari, 2026). Our audit links comparisons of multimodal scores and fixed-budget error composition to model configurations, output definitions, populations, and aggregate sources.

E Extended Coverage and Output Diagnostics

E.1 RQ1: Evidence Coverage and Output Semantics

E.1.1 Image availability selects a different population.

Of 14,637 test records, 2,232 had an eligible CXR reference and 2,220 had a readable file (Table 5). The remaining 12,417 records comprised 12,405 without a reference and 12 unavailable files. Readable-image records had higher outcome prevalence than records without readable images: mortality differed by 0.0427 (95% bootstrap CI [0.0273, 0.0579]) and LOS by 0.1313 ([0.1093, 0.1523]). This test-partition pattern is consistent with the source benchmark's reported association between multimodal completeness and adverse outcomes (Yu et al., 2025); the audit additionally separates eligible references from readable files. In the 2,247-record complete-case encoder experiment, GRU+RadBERT+Swin achieved the strongest mortality AUROC (0.8227), while GRU+RadBERT achieved the strongest LOS AUROC (0.7935); adding Swin yielded LOS AUROC 0.7865. Across three seeds, AUROC standard deviations were at most 0.0034. In the separate full-cohort coverage-aware experiment, a structured-only logistic model supplied predictions for all records, and a separately fitted multimodal logistic model replaced its predictions for the 2,203 test records with complete features. This routing increased mortality AUROC from 0.7718 to 0.7788 and LOS AUROC from 0.6603 to 0.6748. These within-cohort comparisons show task-dependent modality contributions; the complete-case and full-cohort results also reflect different case mixes.

E.1.2 Hard decisions compress variation in model outputs.

Several vision-language protocols concentrated their predictions in one class (Table 6). For mortality, Qwen predicted no positive cases, and Luna and Sol predicted almost none. For LOS, the LLaVA-Med forced-choice rule predicted every case positive. Their class-conditional metrics expose this concentration directly. Among the 2,220 readable-image records, replacing the image with a blank reduced LLaVA-Med mortality AUROC by 0.0209 (95% CI [0.0002, 0.0410]) and AUPRC by 0.0226 ([0.0070, 0.0399]), while changing only 0.14% of mortality hard answers. The continuous score was sensitive to the image even when the thresholded decision was nearly unchanged. LOS effects were uncertain. Records without readable images received identical placeholders in both image conditions.

Table 6: Hard-decision metrics on 14,637 test records under each specified model and output protocol.

Model/protocol	Outcome	Bal. acc.	Sensitivity	Specificity	Predicted +
LLaVA-v1.5 free generation	Mortality	0.5777	0.4639	0.6915	32.64%
LLaVA-v1.5 free generation	LOS	0.5003	0.0025	0.9981	0.21%
LLaVA-Med forced choice	Mortality	0.5245	0.9610	0.0879	91.77%
LLaVA-Med forced choice	LOS	0.5000	1.0000	0.0000	100.00%
Qwen3.5-9B	Mortality	0.5000	0.0000	1.0000	0.00%
Qwen3.5-9B	LOS	0.5035	0.0099	0.9972	0.51%
GPT-5.6 Luna	Mortality	0.5003	0.0006	1.0000	0.007%
GPT-5.6 Luna	LOS	0.5279	0.0986	0.9573	6.09%
GPT-5.6 Sol	Mortality	0.5006	0.0012	1.0000	0.014%
GPT-5.6 Sol	LOS	0.5400	0.1477	0.9324	9.37%

E.1.3 Score range and parseability leave calibration unresolved.

LLaVA-Med mean candidate scores were 0.5112 for mortality and 0.7775 for LOS, above the observed prevalences. Brier skill relative to the prevalence-only diagnostic was negative for both outcomes, and retaining the 20% largest score margins yielded balanced accuracy at most 0.5. Luna and Sol achieved 100% parse coverage, while their candidate-score coverage ranged across split and outcome from 0.007% to 9.03% and 0.15% to 26.77%, respectively. These protocols expose full-cohort hard decisions and only partially observed graded scores. Coverage, decision behavior, and calibration each describe a different property of the same output pipeline.

F Extended Pipeline and Evaluation Protocol

F.1 Coverage and Output Diagnostics

Coverage is measured separately for an eligible CXR reference and for local file readability. Missing-image handling is protocol-specific: local image pipelines use a placeholder, while the API runner omits unavailable image content. Original-versus-blank comparisons hold the model, text, and prompt fixed and replace the image within the readable-image subset. Encoder reconstruction supplies a complementary analysis of modality combinations in complete-case and full-cohort settings; its representations and cohort definitions appear in Appendix B.

Archived vision-language diagnostics cover LLaVA-v1.5-7B free generation (Liu et al., 2024a), forced-choice LLaVA-Med-v1.5-Mistral-7B (Li et al., 2023), Qwen3.5-9B, and API model identifiers recorded as gpt-5.6-luna and gpt-5.6-sol. For LLaVA-Med, the continuous score is the two-candidate softmax mass for Yes at the next token; its hard decision is Yes when this score is at least 0.5. LLaVA-v1.5, Qwen, Luna, and Sol hard decisions are parsed from generated text, with candidate scores recorded as separate output channels.

Luna and Sol each have 7,318 validation and 14,637 test records per outcome. The archived runner supplies deidentified first-window context and an available routed image at low detail, requests a short Yes/No answer, and sets `store=false`. It scans generated token positions and uses the first position exposing both labels among the top-20 candidates plus the emitted token. After trimming whitespace and ignoring case, it keeps each label's maximum log probability and normalizes their exponentiated masses. Thus candidate scores describe the first qualifying position. We report availability alongside parsed hard decisions; AUROC from a binary prediction equals balanced accuracy. The supplement records the prompt template, request settings, and extraction rule; Appendix B specifies other output diagnostics.

F.2 Expert Scores and Fusion

The gate receives five scalar expert scores: structured vital-sign, demographic, and admission features (p_v), laboratory summaries (p_l), a classifier over GRU representations (p_g), a classifier over RadBERT representations (p_r), and the Luna auxiliary channel (p_a). Images reach the gate through the auxiliary channel. Sol is evaluated as a separate comparator. Clinical experts use median imputation, missing-

cell indicators, standardization, and class-balanced logistic classifiers; large embedding inputs use a log-loss SGD classifier. Base experts are fitted on the training partition.

Naturally unavailable GRU or RadBERT scores use outcome prevalence among that branch's available training records, and the corresponding mask is cleared. The auxiliary input combines a finite candidate score, when available, with a fallback parsed No/Yes answer mapped to $\epsilon/1-\epsilon$, where $\epsilon = 10^{-5}$. Candidate scores are clipped to this interval, and m_s records their availability. This hybrid score is the auxiliary fusion input; standalone hard-decision comparisons use the parsed decisions.

Let m_g and m_r indicate available temporal and report representations, and let m_i be the benchmark image-routing flag. Routing and readable-file coverage are measured separately. With $\ell(p) = \log(p/(1-p))$ after clipping and $\mathbf{p} = (p_v, p_l, p_g, p_r, p_a)$, the gate uses 15 features:

$$\begin{aligned}\mathbf{x} = \big[&\ell(p_v), \ell(p_l), \ell(p_g), m_g, \ell(p_r), m_r, \ell(p_a), m_i, \\ &\operatorname{mean}(\mathbf{p}), \operatorname{sd}(\mathbf{p}), \operatorname{range}(\mathbf{p}), \ell(p_a)m_i, \ell(p_a)m_s, |p_a - p_g|, |p_a - p_l|\big].\end{aligned} \tag{8}$$

For outcome t, the fused score is

$$s_t = \sigma\big(b_t + \mathbf{w}_t^{\mathsf{T}} \operatorname{Standardize}(\mathbf{x})\big)\,. \tag{9}$$

The task-specific coefficients are fixed after fitting; interaction terms express dependence on evidence state. Scores-only fusion S uses just the five clipped logits. The clinical stack uses four clinical logits, GRU/report availability and image-routing flags, and mean, population standard deviation, and range over those four scores. It excludes the auxiliary score.

The clinical stack, S, and CARE-Fuse-U each fit a standardized, class-balanced, L2-regularized logistic classifier on validation. Five-fold stratified cross-validation selects $C \in \{0.001, 0.01, 0.1, 1, 10\}$ using pooled out-of-fold AUROC. The threshold maximizes out-of-fold balanced accuracy over unique score quantiles at 193 positions between 0.02 and 0.98. Each selected gate is refitted on all validation records. All four variants expose fitted sigmoid scores without post-hoc calibration. Class balancing optimizes a reweighted objective, so probability interpretation requires assessment on the original event distribution.

F.3 Perturbation Training and Stress Evaluation

CARE-Fuse-R uses the same feature equation and classifier family with four validation-time training perturbations: GRU unavailability, RadBERT unavailability, joint GRU/report unavailability, and auxiliary conflict. Each condition uniformly selects 10%, 30%, or 50% of all records using three augmentation seeds, including records whose affected branch is already unavailable. The original block and 36 augmented blocks produce 37 variants per source record, including unchanged copies of unselected records.

Removal replaces the affected GRU or RadBERT score with full-training-cohort outcome prevalence and clears its mask. Report-output removal also sets the image-routing feature to zero; the archived auxiliary score remains unchanged. These interventions modify the frozen score-and-flag vector. Auxiliary conflict replaces the auxiliary score with $1 - \operatorname{mean}(p_v, p_l, p_g, p_r)$ and preserves the original candidate-score availability mask. Vital-sign and laboratory scores stay intact. Selection and score alteration use outcome-independent rules; supervised fitting uses the labels.

All variants of a source record stay in one cross-validation fold. Regularization and threshold selection use predictions on the clean held-out validation block, and the final gate fits all 37 blocks. U and R have distinct fitting seeds and select their own regularization and thresholds. Their comparison evaluates these two fitted training variants within the common feature and classifier family.

Test perturbations use five draws with seeds separate from training, applying the same four families and three rates used in augmentation. We report five-draw means for all 12 conditions per outcome and each model's minimum condition mean. Joint temporal/report removal and auxiliary conflict illustrate the contrasting responses. We align S's archived results to the R evaluation by outcome, scenario, rate, and seed, checking shared U/clinical aggregate metrics. Paired bootstrap contrasts available for R versus U and clinical refer separately to the first draw; S contrasts are descriptive. Appendix C gives configuration and seed details.

We also measure each model's maximum gap to the condition-specific best mean AUROC. For outcome t, model k, and condition c, let $A_{t,k,c}$ be the five-draw mean. Define

$$G_{t,k} = \max_{c} \left(\max_{j} A_{t,j,c} - A_{t,k,c} \right). \tag{10}$$

The inner maximum is a retrospective reference over the four models; the outer maximum covers the 12 nonzero stress conditions. This descriptive summary measures relative shortfall separately from the minimum absolute AUROC.

F.4 Fixed-Budget Ranking Policies

RQ3 uses refitted CARE-Fuse-U scores s and threshold τ, with proximity P, disagreement D, and missingness M:

$$P = 1 - \mathrm{clip}\left(\frac{|s - \tau|}{\max(\tau, 1 - \tau, \epsilon)}, 0, 1 \right), \tag{11}$$

$$D = \mathrm{sd}(\mathbf{p}), \tag{12}$$

$$M = (1 - m_g) + (1 - m_r) + (1 - m_s) + (1 - m_i). \tag{13}$$

Let F_P, F_D, F_M denote the empirical CDFs of P, D, M on validation, respectively. Values of P use in-sample scores from the gate refitted on the full validation set. The combined priority is

$$\rho = \big(F_P(P) + F_D(D) + F_M(M)\big)/3. \tag{14}$$

The components of M record representation availability, image routing, and candidate-score availability. P is a deterministic threshold-margin heuristic, named `uncertainty_only` in the source tables.

For each budget $b \in \{0.05, 0.10, 0.20, 0.30\}$, each policy selects the top $\mathrm{round}(bN)$ records by ρ, $F_P(P)$, or $F_D(D)$, respectively. All CDFs use right-inclusive comparisons. These step functions can map distinct raw values to the same priority. The archived implementation reverses default `numpy.argsort`, without an explicit stable tie key; the reported counts describe those selections. Error capture is the selected fraction of all incorrect thresholded predictions; false-negative capture is the selected fraction of all false negatives. Error enrichment divides error capture by the actual selected fraction. Predictions remain fixed throughout selection and retained-set evaluation.

F.5 Error Decomposition and Cost Sensitivity

For cohort size N, total errors E, and false negatives F, the source fields give $E = N\, r_{\mathrm{error}}$ and $F = N\, r_{\mathrm{FN}}$. The field r_{FN} is the full-cohort fraction FN/N; conditional false-negative rate is $\mathrm{FN}/(\mathrm{FN} + \mathrm{TP})$. We recover FP as $E - F$ and obtain TP and TN from the known outcome counts.

Total-error, false-negative, and positive-case capture fractions recover the selected errors and positives for each policy and budget. Their differences yield selected FP and TP, and the selected cohort size supplies TN. Subtracting gives the retained confusion table. Seventy-six floating-to-integer checks agree within 10^{-12} of an integer, and every confusion table sums to its known cohort size.

Taking an incorrect prediction as the positive detection label, total-error capture is error-detection recall. If a policy selects m records containing E_{sel} errors, write $q = m/N$, $e = E/N$, and $c = E_{\mathrm{sel}}/E$. Detection precision is ec/q, and retained ordinary error is $e(1 - c)/(1 - q)$. For the same classifier, cohort, and actual selection fraction, higher capture therefore means higher precision and lower retained ordinary error. Separating the original classifier's FP and FN allows their relative costs to change this policy ordering.

For a policy π selecting m records, define retained weighted error using the false-negative/false-positive cost ratio $\lambda = c_{\mathrm{FN}}/c_{\mathrm{FP}} > 0$ and $c_{\mathrm{FP}} = 1$:

$$L_{\mathrm{ret}}(\pi, b; \lambda) = \frac{\mathrm{FP} - \mathrm{FP}_{\mathrm{sel}}(\pi, b) + \lambda\big(\mathrm{FN} - \mathrm{FN}_{\mathrm{sel}}(\pi, b)\big)}{N - m}. \tag{15}$$

This objective measures weighted errors per retained record and may exceed one. Selected records are excluded from this metric, with their original predictions preserved. At a fixed budget, minimizing

L_{ret} is equivalent to maximizing $\text{FP}_{\text{sel}} + \lambda \text{FN}_{\text{sel}}$. For policies A and B with different selected FN counts, their costs cross at

$$\lambda^* = \frac{\text{FP}_{\text{sel,A}} - \text{FP}_{\text{sel,B}}}{\text{FN}_{\text{sel,B}} - \text{FN}_{\text{sel,A}}}. \tag{16}$$

This is a positive-cost crossover only when the ratio is positive. Policies with equal selected FN counts have parallel cost lines, or coincide when their selected FP counts also agree. Cost-curve analysis compares classifiers across misclassification costs using intersections and lower envelopes (Drummond & Holte, 2006). We apply this comparison to each policy's retained confusion counts, computing intersections and the lower envelope with exact rational arithmetic among the three fixed rankings at each budget. Cost ratios are sensitivity parameters, and the frontiers describe the observed test aggregates. We also calculate retained prevalence, ordinary error, and conditional false-negative rate to characterize the selected case mix.

Using all N inputs as the denominator instead gives generalized weighted risk, $(1 - m/N)L_{\text{ret}}$ (Traub et al., 2024). The two objectives rank policies identically at the same budget. We report the four observed budgets separately; complete ranking curves are not available in the aggregate archive.

F.6 Statistical Analysis

We measure ranking with AUROC and AUPRC, hard decisions with sensitivity, specificity and balanced accuracy, and calibration with Brier score and equal-width ECE. In the main fusion analyses, AUPRC denotes non-interpolated average precision. The four-model clean and stress evaluation uses 15 bins on $[0, 1]$. Appendix C separates the bin definitions of other experiments. The prevalence-only Brier reference $\pi(1 - \pi)$ uses observed test prevalence as a descriptive diagnostic.

Stress contrasts use outcome-stratified, paired, 2,000-draw record bootstraps conditional on the first test perturbation draw. Image-removal contrasts also use 2,000 draws; initial-fusion comparisons in the appendix use 5,000. The intervals are nominal 95% intervals for their specified contrasts. Five-draw stress means summarize variation across perturbation realizations; the first-draw bootstrap estimates record-resampling variation within one realization. Ranking and cost comparisons are descriptive aggregate analyses. Model fitting and threshold selection use training and validation, while successive reconstruction, fusion and mechanism analyses share the fixed test partition. The multiple comparisons are exploratory.

G Additional Interpretation and Component Implications

G.1 Evidence Coverage Defines the Comparison Population

Readable-image records had additional evidence and a different outcome distribution: this minority had higher mortality and prolonged-stay prevalence. The complete-case and full-cohort experiments consequently answer different comparison questions. For multimodal system evaluation, coverage is part of the population definition. Reporting the denominator and outcome distribution alongside a modality comparison makes the population shift visible.

The output diagnostics extend this argument from inputs to predictions. An image change moved LLaVA-Med's mortality ranking while leaving almost all hard decisions unchanged. Luna and Sol exposed parsed decisions much more often than candidate scores. The computational operation determines which aspect of this behavior matters: ranking consumes score variation, a thresholded action consumes class decisions, and probability display depends on calibration. Evaluating these channels separately helps researchers specify what a prediction component supplies to a system.

G.2 Simple Comparators Change the Interpretation of Robustness

Clean performance placed all four variants close together. Under joint output loss, adding S changed the apparent benefit of perturbation training from a large advantage over U to a small observed difference. Under mortality auxiliary conflict, S declined alongside U, and the clinical stack performed best. R had both the highest minimum condition mean and the smallest maximum gap to the condition-specific best. Comparator selection therefore depends on the particular failure condition and on whether consistency across the tested set is the objective.

Table 7: Observed comparisons that inform prediction and case-ranking component evaluation. These are implications of the computational measurements.

Component operation	Required comparison	Evidence and resulting decision
Order cases under changing evidence	Simple and engineered fusion across the specified failures	S approaches R under joint loss; R has the highest worst-condition mean in the tested set. Compare both the target failure and the broader condition set.
Display a probability	Calibration on the original event distribution	Mortality R ECE is 0.2145 despite AUROC 0.8675. A probability display requires a calibrated output channel.
Allocate a selection budget	Selected FP/FN counts and retained confusion table	At 10% LOS selection, the preferred ranking reverses near a cost ratio of 1.57. Specify the error objective before choosing the policy.

The large R-minus-U LOS contrast occurs under coupled score replacement and availability-flag changes, with the original auxiliary score retained. The result therefore characterizes the fusion component's response at that interface. Including the simple transformation and the full condition set helps compare sensitivity to this intervention with performance in other conditions. Training and testing use the same failure families.

Class-balanced fitting also helps interpret the score diagnostics. These models optimize a class-weighted loss, so their raw sigmoid outputs require assessment on the original event distribution. Prior work documents calibration harm from resampling-based imbalance corrections in logistic risk prediction (van den Goorbergh et al., 2022). Our aggregate results identify the calibration of these fitted outputs; separating the contributions of weighting, regularization, and fusion requires a matched calibration comparison.

G.3 A Ranking Policy Encodes an Error Objective

The cost analysis connects two decisions within the same pipeline: choosing a classifier threshold and evaluating a selection policy. For these imbalanced outcomes, balanced-accuracy threshold selection assigns higher relative cost to false negatives, while total-error capture assigns equal cost to both error types. In this archive, carrying the validation-derived weights into the 10% selection comparison favors proximity; equal-error weighting favors combined. Different pipeline stages can use different objectives, and the comparison makes their consequences explicit. Combined generally captures more false positives; threshold proximity captures more false negatives. The crossover expresses this exchange in a form that can be compared across outcomes and budgets. Its much lower value for LOS at 10% selection means that a smaller change in relative error cost reverses the ordering in that task.

Selection also changes the population on which remaining performance is measured. The mortality example shows retained accuracy increasing while sensitivity decreases, driven by the removal of many true positives and few false negatives. A queue or selective-prediction component should therefore be evaluated with both its selected error composition and its retained confusion table. A coverage–error summary gains an operational interpretation when it includes the relative error objective and retained outcome prevalence.

Table 7 connects the findings to concrete component-evaluation decisions. The shared lesson is to make the comparison match the operation: a score can be adequate for ordering yet poorly calibrated, and a ranking can capture more errors while selecting fewer of the errors assigned greater cost.