# TASKEVI-FUSE: TASK-ADAPTIVE CLINICAL EVIDENCE FUSION FOR MULTIMODAL ICU OUTCOME PREDICTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

ICU outcome prediction must integrate structured EHR histories, chest radiographs, and radiology reports, although the contribution of each modality may vary by outcome. We present TaskEvi-Fuse, a multimodal clinical prediction framework that combines Gated Residual Modality Alignment (GRMA), which refines representations through gated residual updates and modality-subset experts, with Validation-Controlled Risk Construction (VCRC) for outcome-specific risk-policy selection and calibration. We evaluate the frozen endpoint-specific primary system separately from a controlled common-bank analysis. On a subject-disjoint MIMIC-IV/MIMIC-CXR cohort, using three initializations on one fixed split, the primary system achieves AUROC/AUPRC values of 0.8438/0.4432 for in-hospital mortality, 0.7419/0.7797 for long-stay, and 0.5895/0.3072 for readmission. Among the evaluated protocol-adapted multimodal comparators, it ranks best across the six reported metrics for both long-stay and readmission, and on four of six mortality metrics. Controlled analyses reveal outcome-dependent patterns: different representation variants and risk policies are favored for different endpoints. These findings highlight the value of evaluating multimodal representation refinement and risk construction across clinical outcomes rather than assuming a uniform effect.

## 1 INTRODUCTION

Intensive care admissions generate structured measurements, chest radiographs, and radiology reports that provide complementary but uneven evidence. A modality can help one outcome and contribute little to another; moreover, combining predictors is only useful if the final probabilities are reliable (Van Calster et al., 2019). Existing clinical multimodal approaches study representation fusion, missingness, or expert routing (Hayat et al., 2022; Yao et al., 2024b; Han et al., 2024). We study the related system-level question of how to construct an outcome-specific risk from heterogeneous candidate predictors, while retaining a representation-level alignment module as one candidate source.

TaskEvi-Fuse combines Gated Residual Modality Alignment (GRMA), which refines multimodal representations and generates an alignment-aware candidate, with Validation-Controlled Risk Construction (VCRC), which applies development-data risk construction to each endpoint. Because the candidate sources and selected risk paths can differ across outcomes, we state the fitting, selection, calibration, and thresholding path used for each primary endpoint rather than implying one identical executable configuration produced every result.

We evaluate the framework at two linked levels. The primary benchmark reports the frozen endpoint-specific system: mortality uses a validation-selected logistic stack over 17 saved risks; long-stay and readmission each use an 11-risk logistic stack fitted and selected on validation, followed by endpoint-specific Platt calibration and thresholding. A separate ten-candidate common-bank analysis examines a shared VCRC procedure. Together, these evaluations describe the system-level endpoint routes and the behavior of risk construction under a shared candidate inventory.

Our contributions are:



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

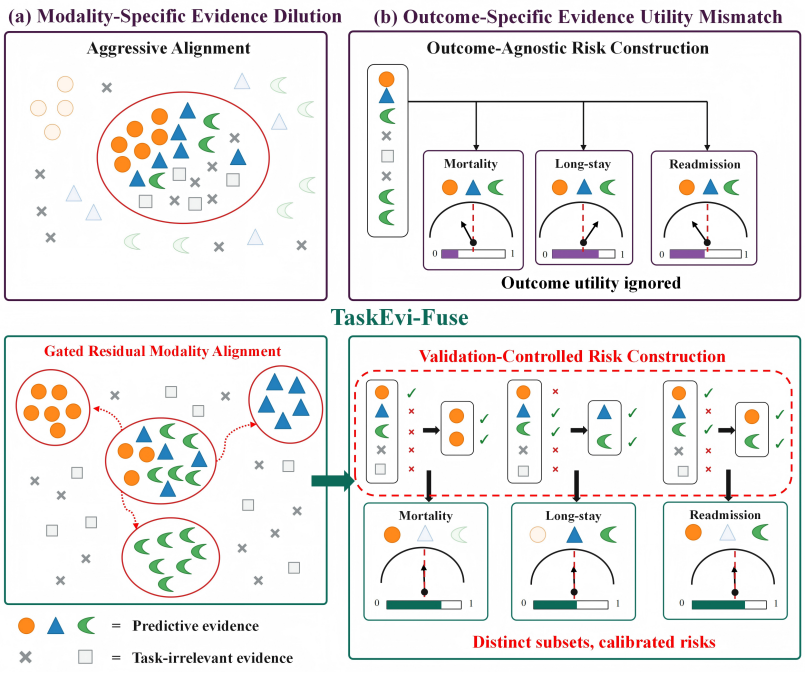


Figure 1: Motivating design questions: multimodal evidence can contain irrelevant signals, and useful candidate evidence can vary by clinical endpoint. The framework therefore separates representation-level candidate generation from endpoint-level risk construction, evaluated through the primary system and a common-bank analysis.

- We formulate TaskEvi-Fuse as a multimodal clinical prediction framework that links gated residual modality refinement to endpoint-specific candidate-risk construction.
- We specify VCRC as a development-data procedure for candidate selection, optional logistic risk stacking, calibration, and operating-threshold choice, and report the fitting and selection data roles for each primary endpoint.
- We evaluate three ICU outcomes with protocol-adapted multimodal comparators and controlled analyses, revealing outcome-dependent prediction profiles and risk-policy choices across mortality, long-stay, and readmission.

**Scope of the contribution.** Gated residual updates, modality-subset experts, logistic stacking, and Platt scaling are established ingredients (Jacobs et al., 1991; Wolpert, 1992; Platt, 2000). We do not claim each as a new algorithm. Our contribution is a clinical prediction framework that connects representation-level candidate generation to endpoint-specific risk construction, evaluated both as a frozen system and under a shared candidate inventory. We do not claim prospective clinical utility, external-site validity, or universal superiority over every comparator.

**Related work and positioning.** Earlier clinical systems combine physiological time series with notes or CXRs (Khadanga et al., 2019; Hayat et al., 2022); tri-modal prediction has also used physiological signals, images, and notes (Wang et al., 2024). Missing or inconsistent modalities and delayed imaging motivate latent imputation, shared-specific representations, and individualized image generation (Zhang et al., 2022; Yao et al., 2024b;a). More general approaches adapt modality weights or expert availability (Gao et al., 2024; Han et al., 2024; Yun et al., 2024). ShaSpec and DANTE additionally model shared and distinctive features or cross-modal transfer (Wang et al., 2023; Zhang et al., 2026). These works provide relevant representation-level alternatives; our available controlled comparison does not establish that GRMA outperforms each one. At the risk level, stacked generalization and Super Learner motivate fitting combinations on out-of-sample predictions (Wolpert, 1992; van der Laan et al., 2007). Calibration and reporting guidance motivate evaluating probability quality separately from discrimination (Guo et al., 2017; Steyerberg et al., 2010; Collins

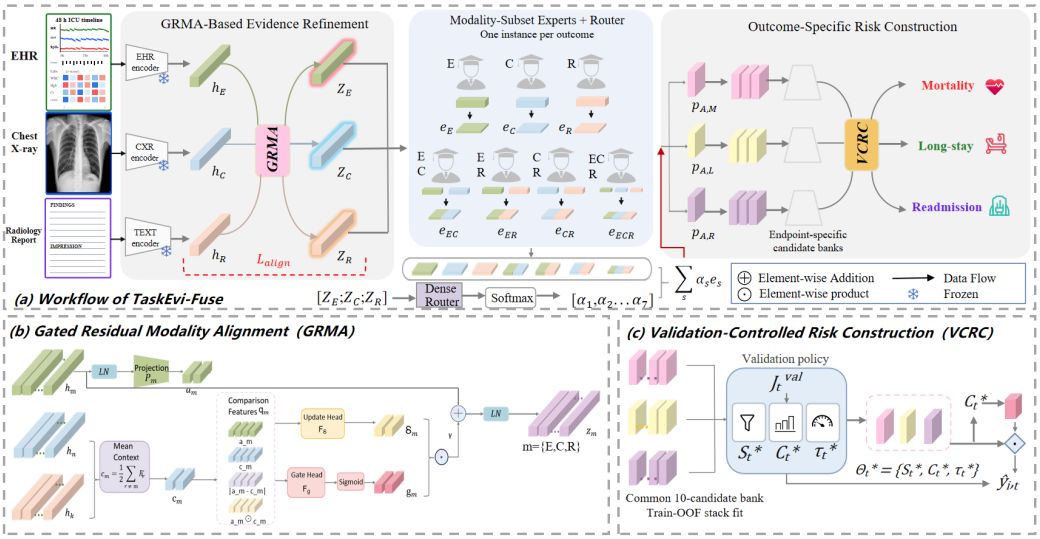


Figure 2: TaskEvi-Fuse framework overview. (a) Multimodal inputs pass through GRMA and outcome-specific modality-subset expert/router instances to form endpoint-specific candidate-risk banks. (b) Details GRMA's gated residual update. (c) Shows the common-bank VCRC analysis used to examine endpoint-specific risk construction. Panel (a) is a framework schematic; the exact primary paths are listed in Table 1. Panel (c) shows the controlled common-bank analysis and is not a reconstruction of the primary predictors.

et al., 2024). Reusing development observations for several selection decisions can yield optimism (Varma & Simon, 2006), which affects all frozen primary routes that reuse validation for fitting or selection.

## 2 METHOD

### 2.1 COHORT AND PREDICTION TARGETS

For admission $i$, the available modalities are structured EHR, chest X-ray (CXR), and radiology report. The prediction landmark is 48 hours after hospital admission. We evaluate post-landmark in-hospital mortality, hospitalization longer than seven days, and a subsequent admission observed in the database within 30 days after discharge alive. Mortality and long-stay contain 7,626/1,638/1,634 train/validation/test admissions; readmission contains 6,789/1,453/1,460. The corresponding test prevalences are 10.65%, 56.36%, and 24.32%.

The principal train, validation, and test partitions are subject-disjoint. Available provenance does not include a complete per-study CXR timestamp audit or final report-signing times at the landmark. Planned readmissions are not explicitly documented as excluded, and follow-up outside the observed health system is unavailable. We therefore describe this as an internally evaluated, database-observed prediction task rather than a fully verified prospective landmark deployment study.

### 2.2 GRMA CANDIDATE GENERATION

GRMA is the representation-level candidate generator studied in the controlled analysis. Let $\mathbf{h}_i^m$ denote the 512-dimensional representation for modality $m \in \{E, C, R\}$. The retained representation-stage description uses recurrent EHR features, ResNet-based CXR features, and Doc2Vec/recurrent report features (Hochreiter & Schmidhuber, 1997; He et al., 2016; Le & Mikolov, 2014); upstream training provenance is incomplete for some archived primary candidate streams. For each modality, GRMA forms a projected self feature, the mean of the other modalities, and an interaction descrip-

| Endpoint | Frozen primary path | Development-data use and provenance |
|---|---|---|
| Mortality | Validation-selected logistic stack over 17 saved risks. | Validation fits the stack and selects the rule; $C = 10/10/3$ across seeds 42/43/44. Platt calibration and max-F1 threshold are also validation-based. |
| Long-stay | Validation-fitted logistic stack over 11 risks, followed by Platt calibration. | Validation fits and selects the stack by AUPRC over individual risks, their mean, and a logistic grid; selected $C$ values are $3/3/10$ across seeds. Platt calibration and the max-F1 threshold are selected on validation. |
| Readmission | Validation-fitted logistic stack over 11 risks, followed by Platt calibration. | Validation fits and selects the stack by AUPRC over individual risks, their mean, and a logistic grid; selected $C$ values are $0.1/0.3/10$ across seeds. Platt calibration and the max-F1 threshold are selected on validation. |

Table 1: Endpoint-level prediction paths and development-data roles for the primary evaluation.

tor:

$$\widetilde{\mathbf{h}}_i^m = \mathrm{LN}(\mathbf{h}_i^m), \qquad \mathbf{a}_i^m = P_m(\widetilde{\mathbf{h}}_i^m), \tag{1}$$

$$\mathbf{c}_i^m = \frac{1}{2}\sum_{n\neq m} \widetilde{\mathbf{h}}_i^n, \qquad \mathbf{u}_i^m = [\mathbf{a}_i^m; \mathbf{c}_i^m; |\mathbf{a}_i^m - \mathbf{c}_i^m|; \mathbf{a}_i^m \odot \mathbf{c}_i^m]. \tag{2}$$

Modality-specific correction and gate networks produce

$$\boldsymbol{\delta}_i^m = F_\delta^m(\mathbf{u}_i^m), \qquad \mathbf{g}_i^m = \sigma(F_g^m(\mathbf{u}_i^m)), \tag{3}$$

$$\mathbf{z}_i^m = \mathrm{LN}(\mathbf{h}_i^m + \gamma\, \mathbf{g}_i^m \odot \boldsymbol{\delta}_i^m), \qquad \gamma = 0.2. \tag{4}$$

The sigmoid gate bounds each gate coefficient, but the correction network output is unconstrained; thus this equation does not imply an absolute bound on the residual magnitude. GRMA uses a same-admission cosine-margin loss with margin 0.3 (Hadsell et al., 2006), seven experts over subsets $E, C, R, EC, ER, CR, ECR$, and dense softmax routing. The resulting probability is one alignment-aware candidate, not the final prediction by itself. The controlled implementation uses cross-entropy plus alignment and batch-mean router-balance penalties, each weighted 0.01.

The primary mortality stream named `ada_alignmoe_v2` is one member of the 17-stream primary candidate pool. The retained artifacts do not establish an exact checkpoint/configuration linkage between this primary candidate and the controlled GRMA implementation. The equations above therefore specify the controlled GRMA analysis only; its settings are not attributed to the other primary candidates.

### 2.3 Endpoint-specific frozen primary configurations

The primary results evaluate the frozen endpoint-specific system recorded as `OURS_AlignMoE_ExpertStack_v4`. Its actual prediction paths differ by endpoint, as summarized in Table 1. The final v4 stage consumes saved candidate probabilities; it is not a single end-to-end raw-input checkpoint and its final risk rule is selected at task/seed level, not per patient.

For mortality, the primary stack uses the frozen 17-risk order and source-script logit transform. Validation selects passthrough versus stack and $C$ using $\mathrm{AUPRC} + 0.20\,\mathrm{F1} - 0.01\,\mathrm{NLL}$; the selected $C$ values are 10, 10, and 3. The historical stack coefficients were not persisted. Validation-reconstructed parameters replay the archived probabilities within validation-derived tolerances (Appendix B), which establishes numerical compatibility but not historical coefficients or train-OOF provenance. For long-stay and readmission, the endpoint-specific procedure uses separate 11-risk banks and selects among individual risks, their mean, and logistic candidates by validation AUPRC. The selected rule keys are $3.0_bal1, 3.0_bal1, 10.0_bal1$ for long-stay and $0.1_bal1, 0.3_bal0, 10.0_bal0$ for readmission. The calibration stage selects positive-slope Platt scaling by validation NLL (AUPRC tie-break), refits on validation, and fixes the max-F1 threshold. Historical stack coefficients and train-OOF/fold-exclusion artifacts are unavailable for these routes. Per-seed rules and calibration/threshold values are summarized in Appendix A. These validation-fitted

primary routes remain descriptive and may retain selection optimism; the controlled common-bank analysis is a separate procedure with OOF-fitted stacks.

### 2.4 COMMON-BANK VCRC ANALYSIS

The common-bank analysis applies the same ten candidates to each outcome and seed: seven modality-subset risks, a meta-risk, a deep tri-modal risk, and a GRMA candidate. VCRC considers GRMA-only, revalidated-meta, best-single, and regularized full-bank logistic-stack policies. Stack coefficients in this analysis are fitted on training out-of-fold predictions; the OOF folds are admission-level, so admissions from one subject can occur in different folds. This differs from the subject-disjoint train/validation/test partition and is detailed in Appendix C.

On validation, policies are selected using $\mathrm{AUPRC} + 0.20\,\mathrm{F1} - 0.01\,\mathrm{NLL}$ with deterministic tie-breaking. The stack grid is the $C$ set above. Identity calibration is compared with positive-slope Platt scaling, using a five-fold validation calibration assessment; the selected calibrator is refitted on validation, and the threshold maximizing validation F1 is fixed before test evaluation. These calibration folds do not cross-fit the full policy search. Policy, calibration, and threshold selection reuse validation observations, leaving selection optimism possible (Varma & Simon, 2006).

In Figure 2, panel (a) presents outcome-specific candidate generation, while panel (c) details VCRC under a shared candidate bank. In panel (c), $S_t^*$, $C_t^*$, and $\tau_t^*$ denote the selected policy, calibrator, and threshold, and $\Theta_t^*$ denotes their joint setting; $C_t^*$ is distinct from the stack's inverse-regularization parameter $C$.

## 3 EXPERIMENTS

### 3.1 PROTOCOL AND COMPARATORS

We link MIMIC-IV v1.0 with MIMIC-CXR-JPG v2.0.0 and associated reports (Johnson et al., 2023; 2019). We report AUROC, AUPRC, thresholded F1, ECE, Brier score, and negative log-likelihood (NLL); AUPRC is particularly informative under imbalance (Saito & Rehmsmeier, 2015). The ECE binning rule for the archived primary outputs is not fully documented. Seeds 42–44 vary initialization on one fixed subject split; the seed standard deviation does not measure split- or site-level uncertainty. Test predictions had been viewed during earlier model development, so the reported comparisons are retrospective and should not be interpreted as a sequestered confirmatory evaluation. Baselines are protocol-adapted AdaCoMed, DPMM, MedPatch, and simplified MoMA (Chen et al., 2025; Chan et al., 2025; Al Jorf & Shamout, 2025; Gao et al., 2026). Their predictions are external to the TaskEvi-Fuse candidate bank, so the comparison is contextual and not a matched component ablation.

### 3.2 PRIMARY RESULTS: FROZEN ENDPOINT-SPECIFIC SYSTEM

Table 2 reports the three-seed mean on the fixed test partition. TaskEvi-Fuse has the best displayed long-stay and readmission values on all six metrics, and the best mortality AUROC, AUPRC, Brier, and NLL. MedPatch has the best displayed mortality F1, and MoMA has the best mortality ECE. The readmission AUROC of 0.5895 remains modest in absolute terms. The results support a favorable internal comparison profile for the frozen endpoint-specific system, not uniform superiority or clinical deployment readiness.

### 3.3 CONTROLLED MECHANISM ANALYSES

The harmonized ten-candidate evaluation tests whether the same risk-selection procedure can make outcome-dependent choices. Table 3 reports the three-seed mean AUPRC/NLL. The controlled selected procedure improves on the best-single rule for long-stay, but it does not beat the fixed full-bank stack on long-stay NLL or readmission AUPRC/NLL. Readmission selects a single candidate in all three seeds, with candidate identity varying across seeds. These results support task-dependent risk construction, not a universally superior aggregator.

| Outcome | Method | AUROC↑ | AUPRC↑ | F1↑ | ECE↓ | Brier↓ | NLL↓ |
|---|---|---|---|---|---|---|---|
| *Mortality* | | | | | | | |
| | TaskEvi-Fuse | **.8438** | **.4432** | .4635 | .0149 | **.0751** | **.2568** |
| | AdaCoMed† | .7126 | .2959 | .3113 | .0279 | .0868 | .3068 |
| | DPMM† | .8389 | .4367 | .4624 | .0160 | .0755 | .2587 |
| | MedPatch† | .8377 | .4417 | **.4753** | .0139 | .0752 | .2582 |
| | MoMA‡ | .8382 | .4382 | .4686 | **.0117** | .0756 | .2588 |
| *Long-stay* | | | | | | | |
| | TaskEvi-Fuse | **.7419** | **.7797** | **.7471** | **.0243** | **.2034** | **.5927** |
| | AdaCoMed† | .7312 | .7725 | .7385 | .0293 | .2081 | .6029 |
| | DPMM† | .7315 | .7758 | .7461 | .0396 | .2085 | .6037 |
| | MedPatch† | .7318 | .7730 | .7450 | .0469 | .2089 | .6046 |
| | MoMA‡ | .7320 | .7725 | .7407 | .0357 | .2078 | .6033 |
| *Readmission* | | | | | | | |
| | TaskEvi-Fuse | **.5895** | **.3072** | **.4074** | **.0151** | **.1804** | **.5436** |
| | AdaCoMed† | .5034 | .2558 | .3861 | .0398 | .1861 | .5598 |
| | DPMM† | .5593 | .2994 | .3785 | .0332 | .1833 | .5531 |
| | MedPatch† | .5629 | .2988 | .3799 | .0409 | .1839 | .5543 |
| | MoMA‡ | .5656 | .2994 | .3983 | .0293 | .1826 | .5511 |

Table 2: Primary results (three-seed mean, one subject-disjoint split). The TaskEvi-Fuse row uses the endpoint-specific configurations in Table 1. † denotes a protocol-adapted implementation and ‡ a simplified protocol-adapted implementation. Boldface marks the best displayed value within each endpoint and metric; the comparator results are contextual and do not constitute matched reproduction of each method.

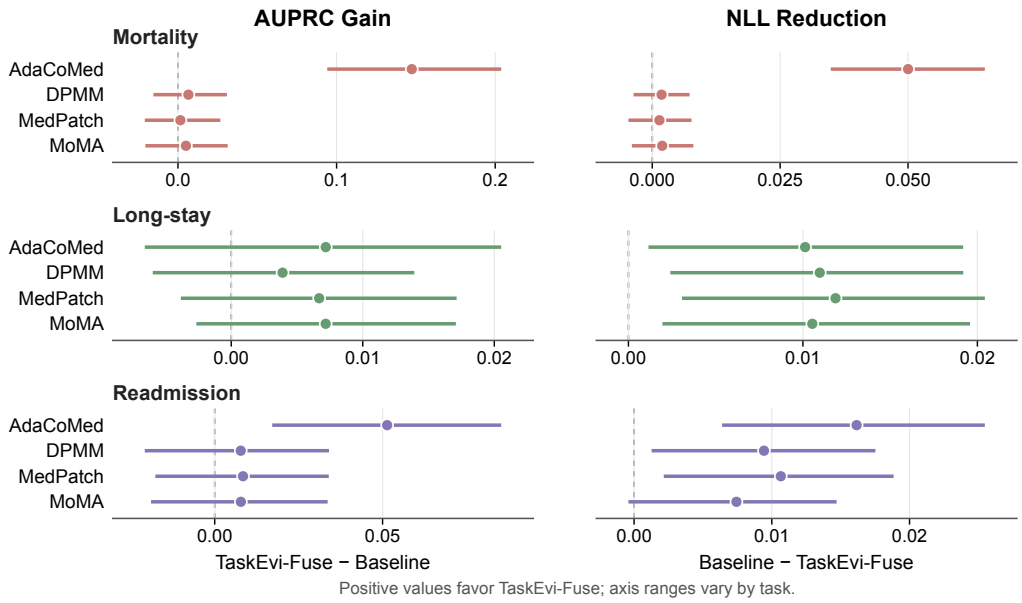


Figure 3: Paired-bootstrap differences between the frozen primary system and protocol-adapted baselines. Points and bars show mean differences and 95% intervals from 2,000 paired subject-level resamples. Positive values favor TaskEvi-Fuse for both AUPRC gain and NLL reduction; axis ranges vary by outcome. The intervals are descriptive; row-level resampling metadata are unavailable for independent regeneration, so we do not claim multiplicity-adjusted significance.

Across the representation ablation, subset experts achieve the highest mean mortality AUPRC, while complete GRMA achieves the highest mean long-stay and readmission AUPRC (Table 4). In a separate active-parameter-matched comparison (6.75M backward-active parameters), the paired subject-

| Policy | Mortality | | Long-stay | | Readmission | |
|---|---|---|---|---|---|---|
| | AUPRC | NLL | AUPRC | NLL | AUPRC | NLL |
| GRMA-only | .4350 | .2609 | .7730 | .6000 | .2954 | .5487 |
| Revalidated meta | .3500 | .2828 | .7171 | .6438 | .3138 | .5457 |
| Best single | .4152 | .2621 | .7730 | .6000 | .3020 | .5488 |
| Fixed full-bank stack | .4085 | .2603 | .7876 | .5884 | .3080 | .5475 |
| Validation-selected VCRC | .4286 | .2604 | .7860 | .5901 | .3020 | .5488 |
| VCRC without GRMA candidate | .3894 | .2622 | .7766 | .5976 | .3030 | .5483 |

Table 3: Separate harmonized ten-candidate controlled analysis. Values are mean test AUPRC/NLL across three seeds. These rows are not the primary Table 2 results. The nine-candidate removal changes candidate availability and is not a parameter- or budget-matched ablation.

| Variant | Active parameters (M) | Mort. AP | Long AP | Readm. AP |
|---|---|---|---|---|
| Concatenation | 1.45 | .4359 | .7686 | .2857 |
| Uniform subset experts | 3.59 | .4431 | .7652 | .2883 |
| Dense router | 3.99 | .4283 | .7697 | .2949 |
| GRMA operator | not recovered | .4366 | .7668 | .2939 |
| Complete GRMA | 6.75 | .4350 | .7730 | .2954 |

Table 4: Controlled representation ablation. AUPRC is averaged across three seeds. The active parameter count for the GRMA-operator-only variant is not available and is not imputed.

cluster bootstrap 95% intervals for full GRMA minus the self-context residual control cross zero for all endpoints: $[-.0171, .0189]$, $[-.0056, .0049]$, and $[-.0027, .0378]$ for mortality, long-stay, and readmission. The ablation therefore shows outcome-dependent component behavior, while the matched comparison does not isolate a reliable AUPRC gain attributable to cross-modal correction alone.

For the same selected controlled risks, validation-fitted positive-slope calibration reduces mean ECE/Brier/NLL from .2518/.1607/.4832 to .0146/.0759/.2604 for mortality, from .0705/.2089/.6071 to .0217/.2027/.5901 for long-stay, and from .2579/.2603/.7316 to .0182/.1817/.5488 for readmission. Because the transformation is monotone, AUPRC is unchanged. This is evidence about the controlled selected risks, not a post-hoc recalibration estimate for the frozen primary outputs.

Figure 5 pairs the representation-level routing pattern with the selected risk-level policies from the controlled common-bank experiment. The policy choices differ across outcomes and seeds; the figure does not depict the archived primary configurations. A three-fold subject-grouped outer validation diagnostic estimates mean development-minus-held-out selection-objective differences of .0055/.0025/.0202 for mortality/long-stay/readmission. It uses smaller development folds than the final policy and cannot exactly correct the optimism in the final validation-selected rules. The auxiliary sensitivity plot in Appendix Figure 6 varies inverse regularization and candidate count; it is not evidence of three-outcome robustness. A local sensitivity analysis over F1 weights $\{0.1, 0.2, 0.3\}$ and NLL weights $\{0.005, 0.01, 0.02\}$ leaves all nine controlled task-seed policy choices unchanged; this demonstrates local stability only.

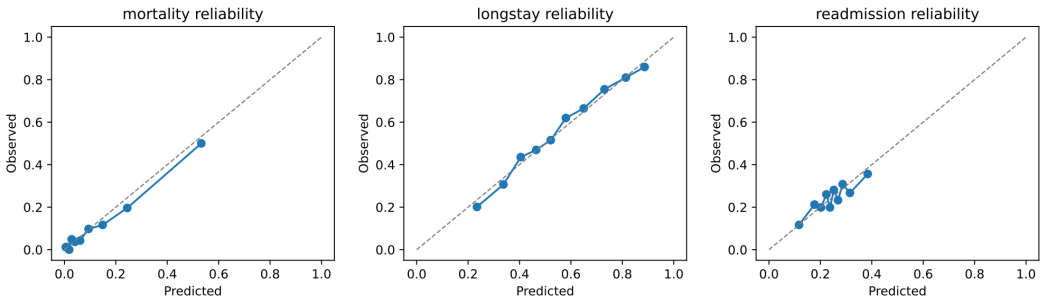


Figure 4: Reliability diagrams for the frozen primary risks across the three outcomes. The diagrams are descriptive calibration summaries; they do not establish external calibration transport.

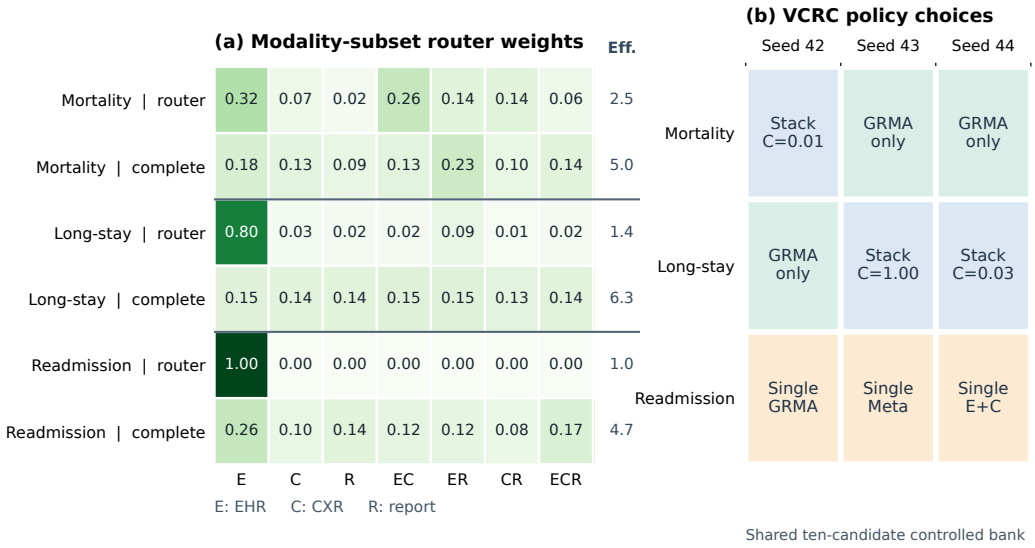


Figure 5: Representation routing and risk-policy choices in the controlled common-bank analysis. (a) Mean weights for the seven EHR (E), CXR (C), and report (R) subset experts in the learned router and complete GRMA variants. Values reproduce the original displayed means rounded to two decimals; "Eff." gives the mean admission-level effective expert count, $\exp[-\sum_s \alpha_s \log \alpha_s]$. (b) Validation-selected VCRC policy for each outcome and initialization seed from the same ten-candidate bank. "Single" identifies the selected individual risk, and $C$ is the inverse regularization of a selected full-bank stack. These selections describe controlled model behavior; they do not represent the distinct frozen primary paths in Table 1 or establish patient-level clinical benefit.

## 4 Discussion and Limitations

The primary evaluation uses endpoint-specific validation-fitted stacks: 17 saved risks for mortality and separate 11-risk banks for long-stay and readmission, each followed by Platt calibration and thresholding. The common-bank analysis separately examines policy behavior under a shared candidate inventory.

The primary table shows a favorable internal comparison profile: TaskEvi-Fuse is strongest on all displayed long-stay and readmission metrics and four mortality metrics among the evaluated protocol-adapted systems. The clinical interpretation must remain proportional to the evidence. Readmission AUROC is 0.5895, baselines are protocol-adapted rather than fully matched reproductions, and no decision-curve or prospective workflow study establishes net benefit to clinicians. Calibration is represented in the primary metrics and reliability plot; same-risk before/after calibration claims are limited to the controlled analysis.

Component performance is outcome-dependent: subset experts have the highest mean mortality AUPRC, while complete GRMA has the highest mean long-stay and readmission AUPRC. Paired intervals against the matched self-context residual control include zero for all three outcomes, so the evidence does not isolate a consistent gain from cross-modal correction. Router weights are descriptive model behavior, not causal evidence that a modality matters to an individual patient.

The evidence is retrospective and limited to one subject-disjoint cohort split: the three seeds measure initialization variability, not cohort or site variation, and test predictions were viewed during earlier model development. For all primary endpoints, validation is used to fit or select the risk stack, calibrator, and threshold. Mortality replay reproduces archived probabilities, but historical stack coefficients and candidate-generator training-fold provenance were not retained. Long-stay and readmission stack coefficients and train-OOF provenance are likewise unavailable. In the common-bank analysis, training OOF folds are admission-level rather than subject-grouped; the grouped outer-validation diagnostic does not correct the final selection process. The available audit also does not fully verify CXR acquisition and report-signing times at the 48-hour landmark, planned-readmission handling, or follow-up outside the observed system. No external multimodal ICU cohort, temporal or prospective validation, common-hardware raw-input latency study, or clinical net-benefit analysis is available. We therefore present TaskEvi-Fuse as a clinically motivated retrospective prediction framework, not a deployable decision-support tool.

## 5 Conclusion

TaskEvi-Fuse combines GRMA-based candidate generation with endpoint-specific risk construction for multimodal ICU prediction. In the primary evaluation, its strongest comparative profile is on long-stay and readmission, while mortality performance varies across metrics. The controlled common-bank analysis shows that preferred risk policies and representation variants differ by outcome. Together, the results support endpoint-specific evaluation of multimodal prediction and calibrated risk construction across ICU outcomes.

### AI use statement

Generative AI tools assisted with experimental planning, analysis-workflow and code review, interpretation of retained results, and manuscript drafting, editing, and conversion to the ICLR format. The authors checked the reported claims and values against the available scripts, logs, prediction artifacts, and result records, and take responsibility for the methods, code, results, figures, and final text.

### Ethics statement

This retrospective study analyzes de-identified public clinical data under their access and use requirements. Predicted risks were not used to direct patient care. Residual data-availability, selection, and transport limitations are described above.

### Reproducibility statement

Appendices A–C report endpoint-level configurations, candidate families, split sizes, seed-level results, policy rules, available OOF-overlap diagnostics, and mortality replay errors. Appendix A summarizes the nine endpoint-by-seed primary configurations, including candidate banks, selected rules, calibration parameters, and thresholds. All three endpoint stacks and subsequent calibration/threshold choices use validation. Historical stack coefficients and candidate-generator train-OOF/fold-exclusion evidence were not recovered. For mortality, validation-reconstructed parameters reproduce the archived probability vectors within validation-derived tolerances (Appendix B); the replay did not use test labels or metrics. Per-seed metrics are transcribed from the archived supplement table rather than recomputed from predictions. A complete per-study timestamp audit is also unavailable, and retained metadata do not identify the exact MIMIC-IV-Note release. These limits prevent exact reconstruction of every upstream primary candidate from the manuscript alone. MIMIC source data remain subject to their original access and credential requirements.

## REFERENCES


B. 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*. PMLR, 2025.

T. H. Chan, F. Wu, Y. Chen, G. Yin, and L. Yu. Amplifying prominent representations in multimodal learning via variational dirichlet process. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

W. Chen, Z. Zhao, J. Yao, Y. Zhang, J. Bu, and H. Wang. Multi-modal medical diagnosis via large-small model collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30763–30773, 2025.

Gary S. Collins, Karel G. M. Moons, Paula Dhiman, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024.

J. Gao, M. Rahman, J. Caskey, M. Oguss, A. O'Rourke, R. Brown, A. Stey, A. Mayampurath, M. M. Churpek, G. Chen, and M. Afshar. Moma: A mixture-of-multimodal-agents architecture for enhancing clinical prediction modelling. *npj Digital Medicine*, 9:46, 2026.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26876–26885, 2024.

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.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 1735–1742, 2006.

Xue Han, H. Nguyen, C. Harris, N. Ho, and S. Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 67850–67900, 2024.

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 the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 479–503. PMLR, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chao-yu 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.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

S. Khadanga, K. Aggarwal, S. Joty, and J. Srivastava. Using clinical notes with time series data for ICU management. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 6432–6437, 2019.

Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1188–1196, 2014.

John C. Platt. Probabilities for SV machines. In *Advances in Large-Margin Classifiers*, pp. 61–74. MIT Press, 2000.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015.

Ewout W. Steyerberg, Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattan. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, 21(1):128–138, 2010.

Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, and Ewout W. Steyerberg. Calibration: The achilles heel of predictive analytics. *BMC Medicine*, 17(1):230, 2019.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):25, 2007.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91, 2006.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multimodal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023.

Y. Wang, C. Yin, and P. Zhang. Multimodal risk prediction with physiological signals, medical images and clinical notes. *Heliyon*, 10(5):e26772, 2024.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

Wenfang Yao, Chen Liu, Kejing Yin, William K. Cheung, and Jing Qin. Addressing asynchronicity in clinical multimodal fusion via individualized chest x-ray generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 29001–29028, 2024a.

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, 2024b.

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-MoE: Modeling arbitrary modality combination via the flexible mixture-of-experts. In *Advances in Neural Information Processing Systems*, volume 37, pp. 98782–98805, 2024.

C. Zhang, X. Chu, L. Ma, Y. Zhu, Y. Wang, J. Wang, and J. Zhao. M3 Care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2418–2428, 2022.

Zhiyun Zhang, Yan-Jie Zhou, Yujian Hu, Xiyao Ma, Zhouhang Yuan, Zirui Wang, Hongkun Zhang, and Minfeng Xu. Disentangling for transfer: Boosting limited modalities via information-theoretic regularization and cross-modal reconstruction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(15):13052–13060, 2026. doi: 10.1609/aaai.v40i15.38305.

# A PRIMARY RESULT DETAILS

| Outcome | Seed | Primary rule | Calibration and threshold | Lineage status |
|---|---|---|---|---|
| Mortality | 42 | 17-risk stack; $C = 10$ | Platt ($a = 1.0026, b = 0.0038$); $\tau = 0.2356$ | Historical stack weights and train-OOF provenance not recovered. |
| Mortality | 43 | 17-risk stack; $C = 10$ | Platt ($a = 1.0018, b = 0.0026$); $\tau = 0.2842$ | Historical stack weights and train-OOF provenance not recovered. |
| Mortality | 44 | 17-risk stack; $C = 3$ | Platt ($a = 1.0023, b = 0.0031$); $\tau = 0.1847$ | Historical stack weights and train-OOF provenance not recovered. |
| Long-stay | 42 | 11-risk; key $3.0_bal1$ | Platt ($a = 0.9962, b = 0.2456$); $\tau = 0.3330$ | Historical stack weights and train-OOF provenance not recovered. |
| Long-stay | 43 | 11-risk; key $3.0_bal1$ | Platt ($a = 0.9955, b = 0.2450$); $\tau = 0.3119$ | Historical stack weights and train-OOF provenance not recovered. |
| Long-stay | 44 | 11-risk; key $10.0_bal1$ | Platt ($a = 0.9972, b = 0.2453$); $\tau = 0.3461$ | Historical stack weights and train-OOF provenance not recovered. |
| Readmission | 42 | 11-risk; key $0.1_bal1$ | Platt ($a = 1.0358, b = -1.1112$); $\tau = 0.1665$ | Historical stack weights and train-OOF provenance not recovered. |
| Readmission | 43 | 11-risk; key $0.3_bal0$ | Platt ($a = 1.0377, b = 0.0403$); $\tau = 0.1639$ | Historical stack weights and train-OOF provenance not recovered. |
| Readmission | 44 | 11-risk; key $10.0_bal0$ | Platt ($a = 1.0052, b = 0.0051$); $\tau = 0.1938$ | Historical stack weights and train-OOF provenance not recovered. |

**Primary configuration summary.** This summary distinguishes the primary routes from the separate common-bank analysis. Each seed is an initialization repeat on one fixed subject-disjoint split. The table above reports the selected rule, calibration parameters, and threshold for each endpoint and seed; the candidate registries below give the available ordered candidate banks. Candidate checkpoint-to-seed/fold mappings, historical stack coefficients, and train-OOF exclusion evidence were not recovered. The per-seed metrics are transcribed from the archived supplement table and were not recomputed from predictions.

| Outcome | Seed | AUROC | AUPRC | F1 | ECE | Brier | NLL |
|---|---|---|---|---|---|---|---|
| Mortality | 42 | .8435 | .4468 | .4808 | .0166 | .0747 | .2562 |
| Mortality | 43 | .8413 | .4398 | .4565 | .0138 | .0758 | .2584 |
| Mortality | 44 | .8465 | .4430 | .4531 | .0142 | .0750 | .2559 |
| Long-stay | 42 | .7410 | .7790 | .7501 | .0221 | .2038 | .5935 |
| Long-stay | 43 | .7443 | .7804 | .7466 | .0316 | .2025 | .5907 |
| Long-stay | 44 | .7405 | .7796 | .7447 | .0192 | .2040 | .5940 |
| Readmission | 42 | .5874 | .3062 | .4046 | .0148 | .1807 | .5448 |
| Readmission | 43 | .5878 | .3144 | .4100 | .0163 | .1799 | .5425 |
| Readmission | 44 | .5932 | .3010 | .4075 | .0143 | .1805 | .5436 |

Seeds 42–44 are initialization repeats on the same subject split. They are not independent cohort replications.

**Candidate registry for long-stay and readmission.** Both endpoints use an 11-risk ordered bank: `image_probe_ehr_imgcxr`, `image_probe_ehr_oldcxr`, `image_probe_ehr_report`, `image_probe_ehr_report_imgcxr`, `step25_alignmoe_imageaware`, `step25_alignmoe_raw`, `step25_concat`, `step25_raw_router`, `step29_task_specific`, `step31_aligned`, and `step5_tri_carr`. Candidate predictions and selected rules are endpoint- and seed-specific. Per-seed calibration and threshold

values are reported in Appendix A. The candidate-to-checkpoint seed/fold mapping and training-OOF exclusion artifacts were not recovered, so the bank is identifiable by its saved prediction artifacts but not fully reconstructible from raw-input checkpoints.

**Candidate registry for primary mortality.** The 17 saved mortality risk streams are `ada_alignmoe_v2`, `biomedcoop_repr`, `ehr_report_concat`, five EHR/report/CXR candidates with `posw` $\in \{3, 4, 5, 6, 8\}$ and dropout settings `0p3, 0p3, 0p3, 0p2, 0p2`, five EHR/report candidates with the same positive-weight/dropout sequence, `step32_expert_stack`, `tri_carr`, `v2_calibrated`, and `v3`. The saved validation and test predictions were joined on subject and admission identifiers. The exact training-fold provenance for all 17 streams and the upstream optimizer/environment record for every generator are unavailable, limiting exact reconstruction of the primary mortality stack from this manuscript alone. The controlled GRMA implementation does not fill those gaps.

## B MORTALITY CONTROLLER REPLAY

The archived mortality run did not retain its 17-input stack coefficient vectors. We reconstructed the parameters separately for each seed from the saved validation predictions in the frozen candidate order, with the source-script logit transform and the recorded regularization setting. We then applied these validation-reconstructed parameters and the saved Platt calibration to the archived test candidate probabilities. This replay used identifiers and probability columns only. Test labels and test metrics were not accessed, and no parameter or decision rule was changed. Row counts and identifier order matched exactly for all 1,634 admissions per seed.

| Seed | $C$ | Validation max error | Test max error | Validation-derived tolerance |
|---|---|---|---|---|
| 42 | 10 | $2.48 \times 10^{-6}$ | $2.16 \times 10^{-6}$ | $4.97 \times 10^{-6}$ |
| 43 | 10 | $2.37 \times 10^{-8}$ | $1.52 \times 10^{-8}$ | $1.00 \times 10^{-7}$ |
| 44 | 3 | $1.28 \times 10^{-11}$ | $1.21 \times 10^{-11}$ | $1.00 \times 10^{-7}$ |

The archived test probabilities are numerically compatible with the validation-reconstructed mortality-stack computation under the tolerances defined from validation replay. This is a prediction-artifact consistency check. It does not recover historical coefficients, establish train-OOF generation for the 17 candidate streams, or turn the already inspected test results into independent confirmation. The candidate order, replay errors, and tolerances are summarized in this appendix.

## C CONTROLLED-ANALYSIS DETAIL

In the harmonized controlled analysis, the ten candidates are seven subset risks ($E, C, R, EC, ER, CR, ECR$), a meta-risk, a deep tri-modal risk, and one GRMA risk. The ten-candidate full stack and reduced nine-candidate diagnostic are different candidate sets; their comparison is not an equal-budget estimate of GRMA's causal effect.

The controlled selected policies by seed are:

| Outcome | Seed | Selected policy | Selected candidate / $C$ |
|---|---|---|---|
| Mortality | 42 | Full-bank stack | $C = 0.01$ |
| Mortality | 43 | GRMA only | – |
| Mortality | 44 | GRMA only | – |
| Long-stay | 42 | GRMA only | – |
| Long-stay | 43 | Full-bank stack | $C = 1.00$ |
| Long-stay | 44 | Full-bank stack | $C = 0.03$ |
| Readmission | 42 | Best single | GRMA |
| Readmission | 43 | Best single | Meta |
| Readmission | 44 | Best single | EHR+CXR |

The training OOF generation for the controlled GRMA candidate is admission-level. The number of training subjects appearing across multiple OOF sides was 638/623/615 for mortality

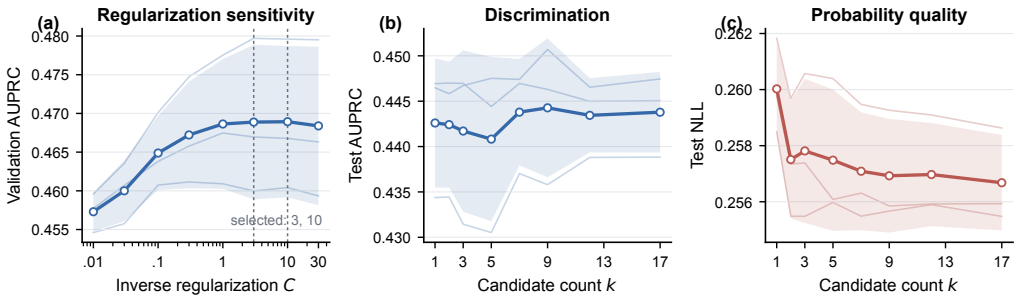


Figure 6: Auxiliary mortality risk-layer sensitivity: (a) validation AUPRC over inverse-regularization $C$, (b) test AUPRC, and (c) test NLL across candidate-bank sizes $k$. Only the displayed configurations are shown; this analysis does not establish stability across outcomes.

seeds 42/43/44; 619/615/608 for long-stay; and 522/512/524 for readmission. These overlaps are within the training fold construction; they do not imply subject overlap across the principal train/validation/test partitions. The grouped outer validation diagnostic is separate and does not alter the reported OOF construction.

For controlled selected-risk calibration, raw versus calibrated test ECE/Brier/NLL are:

| Endpoint | Raw | Calibrated |
|---|---|---|
| Mortality | .2518/.1607/.4832 | .0146/.0759/.2604 |
| Long-stay | .0705/.2089/.6071 | .0217/.2027/.5901 |
| Readmission | .2579/.2603/.7316 | .0182/.1817/.5488 |

The rows use the same frozen selected controlled risk before and after monotone calibration. They do not provide raw-primary predictions for Table 2.

## D Implementation and Interpretation Notes

The frozen primary system is a saved-risk workflow, not an end-to-end model that reruns every encoder at inference. Mortality stack coefficients and the final policy are task/seed-level; the patient-conditioned router exists inside the alignment-aware candidate model only. Figure 5 summarizes average expert weights and does not identify patient-specific causal contribution.

The paired-bootstrap figure reports 2,000 subject-level paired resamples. The exact row-level resampling metadata were not recovered. We therefore use the plot as a descriptive uncertainty summary only and do not infer multiplicity-adjusted significance from it. Figure 6 is likewise limited to the displayed sensitivity range.