Active Learning as Decision Support for EHR Cohort Construction Under Class Imbalance

Eloïse Berson [1 2 3]

Abstract

Building disease risk prediction models from electronic health records (EHRs) requires deciding which patients to include in the training cohort, a decision that shapes predictive accuracy, fairness, and clinical utility. We formulate cohort construction as a sequential decision problem and evaluate unsupervised active learning (AL) strategies as patient selection policies. We benchmark five canonical AL strategies for cold-start EHR disease risk prediction using MOTOR foundation model embeddings, across 18 disease tasks on MIMIC-IV with three classifiers. Entropy sampling is the only strategy that consistently outperforms random across all classifiers, with gains largest for low-prevalence diseases. Diversity-based strategies perform similarly or worse than random because they systematically under-enroll positive-class patients in imbalanced cohorts, regardless of class separability in the embedding space or choice of classifier.

1. Introduction

Building a disease risk prediction model from EHRs requires deciding which patients to include in the training cohort. In prospective biomedical studies, this decision is shaped by strict budgets, clinical capacity, and the competing goal of deliberately over-sampling rare conditions. Cohorts skewed toward majority demographics produce models that fail silently on under-represented groups (Obermeyer et al., 2019; Chen et al., 2021). This has led to rethinking training data composition, prioritizing data quality over data quantity (Zha et al., 2023; Sorscher et al., 2022). AL formalizes this as a sequential selection problem where a strategy selects patients under a limited enrollment budget so that the resulting cohort is maximally informative. Unlike post-hoc data pruning (Sorscher et al., 2022; Mirzasoleiman et al., 2020), AL operates before data are collected, making it directly applicable to prospective study design. Existing AL benchmarks cover single tasks or models (Sener & Savarese, 2018; Ash et al., 2020), focus on imaging or NLP (Gal et al., 2017), or use balanced benchmarks that do not reflect clinical class-imbalance. To our knowledge, no prior study has quantified how AL strategy choice affects prospective patient selection in low-prevalence settings such as EHR-based disease risk prediction.

Our contributions are threefold. (1) We present a reproducible benchmark across 18 tasks (prevalence 1.3%–82.9%), five AL strategies, three classifiers, and three clinical metrics. (2) We establish entropy sampling as the only strategy that consistently outperforms random, with gains largest for low-prevalence diseases. (3) We identify the mechanism whereby diversity strategies systematically under-enroll positive cases, while entropy targets the decision boundary regardless of class separability or classifier type.

2. Related Work

2.1. Training Set Composition

Training set composition, rather than model capacity, is the primary bottleneck in many practical regimes (Zha et al., 2023). While data pruning improves generalization by removing redundant or harmful examples (Sorscher et al., 2022; Mirzasoleiman et al., 2020), it requires access to the fully collected dataset, making it inapplicable to prospective cohort construction. A solution is to select informative examples. For instance, Mindermann et al. (2022) demonstrate that focusing on points that are simultaneously learnable, worth learning, and not yet learned requires $18\times$ fewer training steps. Furthermore, applying informativeness criteria over pre-trained embeddings can drastically reduce data constraints. SemDeDup halves dataset sizes by removing semantic redundancies (Abbas et al., 2023), while similar recent approaches cut data budgets by 70% (Agarwal et al., 2025) and match full-data performance with $13\times$ fewer

[1]Mines Paris, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France [2]Institut Curie, PSL Research University, F-75005 Paris, France [3]INSERM, U1331, F-75005 Paris, France. Correspondence to: Eloïse Berson <eloise.berson@minesparis.psl.eu>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

iterations (Lin et al., 2024; Evans et al., 2024).

2.2. Active Learning

Active learning (AL) formalizes sequential data selection under a labeling budget (Settles, 2009). Three principal strategy families have emerged: uncertainty sampling, query-by-committee, and information-theoretic methods. Early approaches focused on single signals, estimating predictive uncertainty via Monte Carlo dropout (Gal et al., 2017), selecting points nearest the decision boundary (Tong & Koller, 2001), or maximizing spatial coverage via Coresets (Sener & Savarese, 2018). More recent methods attempt to fuse these signals: BADGE (Ash et al., 2020) captures both uncertainty and diversity through gradient embeddings, while diverse mini-batch AL (Zhdanov, 2019) combines both signals via K-means partitioning. Hacohen et al. (2022) establish a phase-transition theory where typical examples are best at low budgets and uncertain examples at large budgets. TypiClust, introduced by the same authors, is designed for the low-budget regime. SelectAL (Hacohen & Weinshall, 2023) dynamically switches between typicality and uncertainty strategies. Class imbalance, the norm in clinical EHR data, is not addressed by these benchmarks, none of which evaluate strategies in imbalanced prospective cohort settings.

2.3. Foundation Models and Data Efficiency

Pre-trained representations substantially reduce the task-specific data needed for clinical prediction. The EHRSHOT benchmark (Wornow et al., 2023) shows that a model pre-trained on 2.57M patient records generalizes to 42 unseen clinical tasks with only a handful of labeled examples. EHR foundation models span health-system language models (Jiang et al., 2023), generalist medical AI (Moor et al., 2023), and structured-record encoders (Rasmy et al., 2021). MOTOR (Steinberg et al., 2021) encodes a patient's full structured history into a fixed-length embedding capturing population-level clinical structure, and remains the only open-source EHR foundation model to our knowledge. Tamkin et al. (2022) show that pre-trained models require up to $5\times$ fewer labels when combined with uncertainty-based AL; Vysogorets & Gopal (2024) demonstrate that frozen pre-trained features are sufficient for efficient AL in text classification. Lu et al. (2025) confirm that uncertainty sampling retains its edge over diversity strategies on tabular datasets when model and query strategy are compatible.

3. Benchmark

3.1. Data Representation and Tasks

We use MIMIC-IV (Johnson et al., 2023) and define 18 binary disease diagnosis tasks spanning cancers, cardiometabolic diseases, and acute conditions (Appendix A), covering a prevalence range from lymphoma (1.3%) to hypertension (82.9%). This work builds upon prior work showing that pre-trained representations substantially reduce annotation cost in AL (Tamkin et al., 2022; Vysogorets & Gopal, 2024). Each patient is represented by a 768-dimensional embedding from MOTOR (Steinberg et al., 2021), a time-to-event foundation model pre-trained on large-scale longitudinal EHR data. Each task is a binary classification at admission. The label is whether a patient has the disease at admission t^*. For cases, t^* is the date of the first qualifying-diagnosis admission. Controls are matched patients admitted without the qualifying diagnosis. MOTOR encodes only events strictly before a feature cutoff $t_f^* = t^* - 90$ days. A 90-day gap excludes the qualifying admission itself and any pre-diagnostic workup, preventing leakage. The full criteria and control-matching procedures are in Appendix B. Both cases and controls are required to have at least one prior admission before t_f^*, ensuring MOTOR encodes a meaningful clinical history for every patient.

We simulate cold-start enrollment. At each round, the AL strategy selects $B{=}2$ patients from the unlabeled pool $\mathcal{U}$ and adds them to the labeled pool $\mathcal{L}$ (i.e., the cohort), and the classifier is retrained from scratch on the updated cohort $\mathcal{L}$. We initialize with $|\mathcal{L}_0|{=}20$ stratified patients and run 490 rounds (up to 1 000 patients total), with five-fold cross-validation. We evaluate AUROC, AUPRC normalized by prevalence ($\text{AUPRC}_\text{n}{=}\text{AUPRC}/\pi$ (Davis & Goadrich, 2006)) to make values comparable across the $64\times$ prevalence range of our tasks, and Brier score (Brier, 1950) at every round and summarize these metrics across rounds using the Area Under the Learning Curve (AULC) as the primary summary statistic (see Appendix B).

3.2. Query Strategies

We evaluate five strategies in total: the uniform random baseline (Random), three model-free strategies that select based solely on embedding geometry and require no labeled data (Coreset (Sener & Savarese, 2018), K-Means (Zhdanov, 2019), TypiClust (Hacohen et al., 2022)), and one model-based strategy that scores candidates using classifier predictions (Entropy). We also include two label-aware enrichment baselines that use ground-truth labels to over-sample positives: a prevalence-stratified Oracle (balanced 50/50 cohorts) and a Biased Random sampler (draws positives with probability $1{-}\pi$, negatives with π, where π is background prevalence). These two baselines isolate how much of entropy's advantage could be attributed to positive-class enrichment alone, providing a reference against which entropy's label-free selection can be benchmarked. The Oracle reflects the clinically standard case-control alternative for rare-disease cohorts, while Biased Random directly tests the

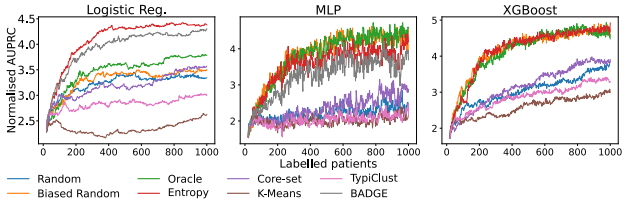


Figure 1. AUPRC_n learning curves across 18 tasks (MOTOR).

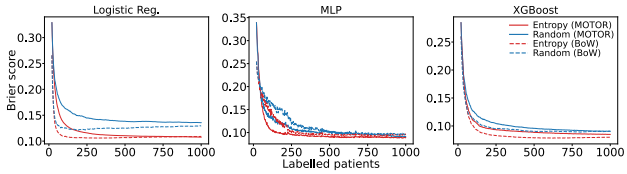


Figure 2. Brier score learning curves, entropy vs. random on MOTOR (solid) and BoW (dashed).

enrichment mechanism without any uncertainty or diversity criterion. Gradient-based methods such as BADGE (Ash et al., 2020) require back-propagation through the classifier and are therefore incompatible with XGBoost, which would prevent fair cross-classifier comparison. We include a supplementary BADGE experiment for logistic regression and MLP only, in Appendix E. We evaluate three classifiers: logistic regression, MLP (two hidden layers), and XGBoost (Chen & Guestrin, 2016), to ensure that our results are robust across different model architectures. Full hyperparameters are in Appendix B.

4. Results

Among unsupervised AL strategies, entropy is the only one that significantly outperforms random across all three classifiers, while all diversity-based strategies are neutral or harmful (see Table 1, Figure 1). The label-aware Biased Random baseline achieves higher raw AUROC and AUPRC_n than entropy for MLP and XGBoost by directly exploiting ground-truth labels; Oracle does the same for AUROC on both classifiers and for AUPRC_n on MLP, but falls slightly below entropy's AUPRC_n for XGBoost. This labeled advantage comes at the cost of calibration: entropy attains the lowest Brier score for every classifier, even compared to these label-cheating baselines. These observations corroborate prior evidence that uncertainty sampling retains its edge over diversity strategies on tabular data (Lu et al., 2025).

Diversity-based strategies perform poorly. K-Means underperforms random by up to 40% relative AUPRC_n AULC on the rarest disease tasks (e.g., leukemia), while Coreset and TypiClust lead to similar performance to random. AUROC confirms the same ranking, with entropy gains $\Delta = +0.010$ to $+0.040$ across classifiers, significantly outperforming random across tasks and classifiers (see Figures 11 and 9).

At the terminal state (Figure 7), entropy exceeds random on 17/18 tasks (mean AUPRC_n 4.50 vs. 3.23), with the largest advantages for the rarest diseases. This confirms that entropy's gains are not an artifact of task difficulty and that patient selection can make a significant contribution to disease risk model performance. Per-task heatmaps and AULC breakdowns are in Figure 10 and Table 3.

5. Ablation

To understand the drivers of this performance gap, we examine four potential mechanisms for the entropy advantage: (1) pre-trained embedding geometry, (2) class separability in the embedding space, (3) positive-class enrichment, and (4) cross-model transfer.

5.1. Input Representation Effects

Given the performance of entropy sampling on MOTOR embeddings, we ask whether this advantage is inherent to the strategy or a product of the representation space. Replacing MOTOR with raw bag-of-words concept-count features, we compare the AULC gain of entropy over random (Δ) under both metrics (Figure 8). In terms of AUPRC_n, entropy's benefits are substantially reduced on BoW, especially for logistic regression. For XGBoost, entropy leads to a larger AUPRC_n gain on BoW than on MOTOR embeddings, with $\Delta{=}{+}1.56$ over random sampling (vs. $\Delta{=}{+}1.14$ on MOTOR). However, in terms of AUROC, entropy falls below random for logistic regression ($\Delta{=}{-}0.037$) and MLP ($\Delta{=}{-}0.016$) on BoW but retains a modest advantage ($\Delta{=}{+}0.008$) with XGBoost. Figure 2 shows the Brier score AULC gain of entropy over random under both feature sets. Entropy improves calibration regardless of feature type. Brier gains are $\Delta{=}{-}0.028$, -0.010, -0.008 on MOTOR and $\Delta{=}{-}0.018$, -0.009, -0.011 on BoW for logistic regression, MLP, and XGBoost respectively. Better-calibrated probability estimates still reward positive-case recovery, which is what AUPRC_n captures, explaining why entropy leads to higher AUPRC_n even on BoW. Entropy selects patients that are uncertain under BoW features, and such patients are not necessarily near the true disease boundary, which might explain why the global AUROC does not improve as it does on MOTOR.

5.2. Diversity Strategies Under-Enroll Positive Cases

While the representation space is a necessary component, it does not fully explain why diversity-based strategies consistently underperform. We therefore investigate the fraction of positive-class patients selected per round relative to background prevalence (enrichment $= 1.0$ means proportional) (Figure 3). K-Means, Coreset, and TypiClust maintain enrichment below 1.0 throughout, systematically starving the

Table 1. AULC summary (MOTOR embeddings, 18 tasks, 5 folds).

	Logistic Reg.			MLP			XGBoost		
Strategy	AUROC	$AUPRC_n$	Brier↓	AUROC	$AUPRC_n$	Brier↓	AUROC	$AUPRC_n$	Brier↓
Random	0.697 ±0.066	3.233 ±2.935	0.146 ±0.086	0.665 ±0.080	2.204 ±1.258	0.109 ±0.059	0.706 ±0.064	3.052 ±2.178	0.103 ±0.069
Entropy	**0.727 ±0.069**	**4.111 ±3.639**	**0.118 ±0.087**	0.706 ±0.079	3.558 ±3.246	**0.099 ±0.067**	0.716 ±0.063	4.192 ±3.943	**0.095 ±0.068**
Coreset	0.700 ±0.066	3.197 ±2.740	0.160 ±0.078	0.667 ±0.082	2.360 ±1.586	0.113 ±0.057	0.707 ±0.066	3.180 ±2.397	0.111 ±0.066
K-Means	0.671 ±0.068	2.353 ±1.545	0.218 ±0.067	0.662 ±0.068	1.999 ±0.907	0.120 ±0.063	0.687 ±0.067	2.583 ±1.542	0.118 ±0.063
TypiClust	0.688 ±0.065	2.863 ±2.293	0.161 ±0.083	0.658 ±0.077	2.037 ±1.050	0.114 ±0.059	0.700 ±0.067	2.877 ±2.055	0.106 ±0.067
Biased Random	0.709 ± 0.066	3.310 ± 2.730	0.228 ± 0.071	0.727 ± 0.061	**3.841 ± 3.368**	0.160 ± 0.064	0.735 ± 0.061	**4.203 ± 3.794**	0.165 ± 0.068
Oracle	0.709 ± 0.067	3.483 ± 2.976	0.239 ± 0.070	**0.728 ± 0.061**	3.823 ± 3.323	0.172 ± 0.063	**0.736 ± 0.061**	4.158 ± 3.485	0.175 ± 0.065

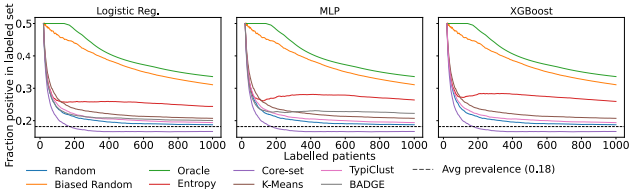


Figure 3. Positive-class enrichment ratio per strategy (18 tasks, 5 folds).

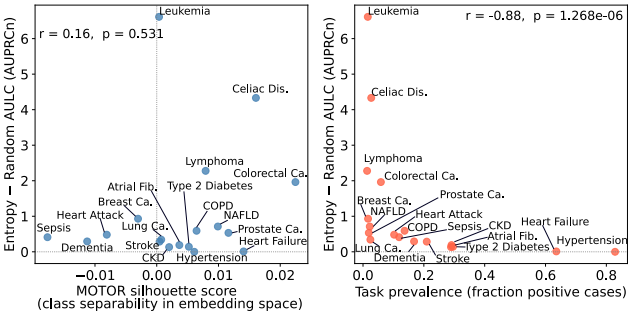


Figure 4. Left: MOTOR silhouette score (class separability) vs. entropy $AUPRC_n$ AULC gain over random, per task. Right: task prevalence vs. entropy $AUPRC_n$ gain.

classifier of rare-event signal. Entropy, however, achieves enrichment at or above 1.0 from round one because it selects boundary samples that are concentrated near the region where positive and negative patients interleave. This effect is strongest for low-prevalence tasks ($\pi < 10\%$), exactly where the $AUPRC_n$ gap is largest, suggesting that positive-class enrichment is a key mechanism for the entropy advantage.

Because Oracle and Biased Random access ground-truth labels from round 0, they exceed entropy's raw AUROC for MLP and XGBoost, and Biased Random also exceeds entropy's $AUPRC_n$ on both classifiers (Oracle falls just short of entropy's $AUPRC_n$ for XGBoost). For logistic regression, however, entropy remains the top performer on every metric (Table 1). Critically, both enrichment baselines substantially degrade Brier score relative to entropy for every classifier, because deliberately skewing training prevalence miscalibrates probability estimates. Entropy, which uses no label information for patient selection, is the only strategy that is simultaneously competitive on ranking metrics and best on calibration. This dissociates two effects: (i) having more positives in the labeled set can help raw ranking metrics for sufficiently flexible classifiers (MLP, XGBoost) but uniformly hurts calibration; (ii) entropy achieves this favorable ranking/calibration trade-off without any label access, because the dominant signal comes from where on the decision boundary patients are selected.

5.3. Entropy Is Robust to Embedding Separability

We then explore whether class separability in the embedding space modulates strategy performance. For this, we compute the silhouette score of the MOTOR embeddings per task, which ranges from -0.018 (least separable) to 0.023 (most separable). We find no correlation between class separability and entropy's $AUPRC_n$ AULC gain (Spearman $r{=}0.16$, $p{=}0.53$; Figure 4, left). In contrast, task prevalence is a strong, significant predictor of entropy's gain (Spearman $r{=}{-}0.88$, $p{<}0.001$; Figure 4, right): the rarer the task, the larger entropy's $AUPRC_n$ advantage over random. Entropy targets the decision boundary regardless of embedding geometry, but its rare-class ranking benefit compounds as background prevalence shrinks, while diversity strategies systematically under-enroll positive cases in imbalanced cohorts.

5.4. Cross-Model Transfer

A potential concern is that entropy's advantage is an artifact of the query classifier. A classifier might select certain patients due to its specific inductive biases rather than their inherent informativeness. To test this, we run a cross-model transfer experiment. More precisely, at each AL round, we compute entropy scores using a logistic regression selector, select patients accordingly, but then evaluate the resulting cohort by training a separately fitted MLP from scratch on the same labeled set. We compare four conditions: (i) Random + MLP, (ii) Entropy + LR, (iii) Entropy + MLP, and (iv) Cross-Entropy (LR selector → MLP evaluator). Conditions (ii) and (iii) are baselines, where the same classifier both selects and evaluates the cohort.

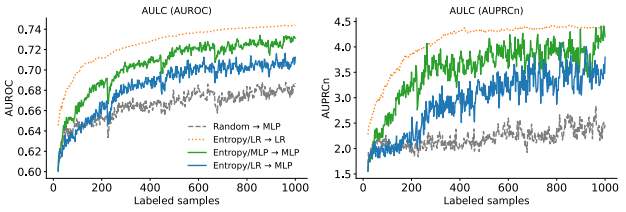


Figure 5. Cross-model transfer: AUPRC_n learning curves (18 tasks, shaded = ± 1 std).

The cross-entropy condition (iv) significantly outperforms random + MLP (i) on average (AULC AUPRC_n: 2.95 vs. 2.20; Wilcoxon $\Delta = +0.74$, $p < 0.001$), beating it on 72% of task×fold pairs and on 13/18 tasks overall (Figure 5). The effect, however, is weaker and less uniform than the matched-classifier baselines. Cross-entropy recovers a median of 47% of the baseline entropy + MLP gain over random (mean transfer fraction 0.52), indicating partial but incomplete transfer of entropy's advantage across classifier families.

Part of entropy's positive-class enrichment is driven by the geometry of the MOTOR embedding space, which is shared across classifier families, so patients near the clinical decision boundary remain informative to some degree regardless of which downstream classifier is used. However, the gap between cross-entropy and the baseline entropy + MLP condition shows that classifier-specific calibration of the entropy score also contributes to the increased performance.

6. Discussion

When labels are not available, entropy sampling with MOTOR embeddings emerges as the preferred choice for prospective EHR cohort construction, consistently and significantly outperforming random across all classifiers, metrics, and 18 disease tasks. Diversity methods fail in imbalanced clinical data because they systematically under-enroll positive cases. This work highlights the importance of patient recruitment strategy in biomedical studies and the value of AL as a decision-support tool for cohort construction.

Limitations. Our benchmark simulates prospective cohort construction using the MIMIC-IV EHR system, which has unusually high data completeness relative to many real-world clinical environments. Further work will validate these findings in prospective real-world settings. In practice, patient selection is further constrained by recruitment cost heterogeneity, multi-site variability, missing data, informed consent, and clinician availability. MOTOR is the only open-source structured-EHR foundation model available to us. Extending the benchmark to other encoders (e.g. masked-LM EHR models (Rasmy et al., 2021)) and multi-site cohorts is an important direction for future work.

Impact Statement

This paper advances machine learning for clinical cohort design. Entropy-based active learning enriches rare-disease cases in training data, with potential fairness benefits. We foresee no harmful consequences beyond the general risk of over-reliance on automated patient selection without clinical oversight.

References

Abbas, A., Tirumala, K., Simig, D., Ganguli, S., and Morcos, A. S. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

Agarwal, I., Killamsetty, K., Popa, L., and Danilevsky, M. DELIFT: Data efficient language model instruction fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2025.

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations (ICLR)*, 2020.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M. Ethical machine learning in health care. *Annual Review of Biomedical Data Science*, 4:123–144, 2021.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 785–794, 2016.

Davis, J. and Goadrich, M. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 233–240, 2006.

Evans, T., Parthasarathy, N., Merzic, H., and Hénaff, O. J. Data curation via joint example selection further accelerates multimodal learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Gal, Y., Islam, R., and Ghahramani, Z. Deep bayesian active learning with image data. In *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1183–1192, 2017.

Hacohen, G. and Weinshall, D. How to select which active learning strategy is best suited for your specific problem and budget. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Hacohen, G., Dekel, A., and Weinshall, D. Active learning on a budget: Opposite strategies suit high and low budgets. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8175–8195, 2022.

Jiang, L. Y., Liu, X. C., Nejatian, N. P., et al. Health system-scale language models are all-purpose prediction engines. *Nature*, 619:357–362, 2023.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-W. H., Celi, L. A., and Mark, R. G. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023.

Lin, Z., Gou, Z., Gong, Y., Liu, X., Shen, Y., Xu, R., Lin, C., Yang, Y., Jiao, J., Duan, N., and Chen, W. Rho-1: Not all tokens are what you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Lu, P.-Y., Cheng, Y.-J., Li, C.-L., and Lin, H.-T. An expanded benchmark that rediscovers and affirms the edge of uncertainty sampling for active learning in tabular datasets. *Transactions on Machine Learning Research (TMLR)*, 2025.

Mindermann, S., Brauner, J. M., Razzak, M. T., Sharma, M., Kirsch, A., Xu, W., Höltgen, B., Gomez, A. N., Morisot, A., Farquhar, S., and Gal, Y. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15630–15649, 2022.

Mirzasoleiman, B., Bilmes, J., and Leskovec, J. Coresets for data-efficient neural network training. In *International Conference on Machine Learning (ICML)*, 2020.

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., and Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature*, 616:259–265, 2023.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447–453, 2019.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4:86, 2021.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018.

Settles, B. Active learning literature survey. *Computer Science Technical Reports*, 1648, 2009.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. S. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Steinberg, E., Jung, K., Shah, N. H., and Fries, J. A. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637, 2021.

Tamkin, A., Nguyen, D., Kadakia, S., Lindsey, J., Liang, P., Goodman, N., and Steinhardt, J. Active learning helps pretrained models learn the intended task. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tong, S. and Koller, D. Support vector machine active learning with applications to text classification. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, 2001. doi: 10.1145/1321440.1321461.

Vysogorets, A. and Gopal, A. Towards efficient active learning in NLP via pretrained representations. *arXiv preprint arXiv:2402.15613*, 2024.

Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M. A., Fries, J., and Shah, N. H. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6:135, 2023.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., and Hu, X. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.

Zhdanov, F. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019.

A. Disease Tasks and Prevalences

Table 2. MIMIC-IV disease prediction tasks and mean cohort prevalence.

Disease task	Prevalence
Lymphoma	0.013
Leukemia	0.016
Breast cancer	0.017
Prostate cancer	0.019
NAFLD	0.022
Lung cancer	0.024
Coeliac disease	0.027
Colorectal cancer	0.059
Heart attack	0.102
Sepsis	0.119
COPD	0.137
Dementia	0.168
Stroke	0.210
Chronic kidney disease	0.289
Atrial fibrillation	0.291
Type 2 diabetes	0.296
Heart failure	0.635
Hypertension	0.829

B. Implementation Details

B.1. Prediction time, feature cutoff, and data leakage

We enforce a strict 90-day gap between the feature window and the prediction date to prevent pre-diagnostic workup events from leaking into the feature representation. Concretely, we distinguish two dates per patient: the prediction time t^* (original diagnosis admission date for cases, matched admission date for controls), which serves as the label and cohort-membership anchor; and the feature cutoff $t^*_f = t^* - 90$ days, which is the actual boundary for all features. MOTOR receives only events strictly before t^*_f; the 90-day gap ensures no pre-diagnostic clinical workup leaks into the embedding. Likewise, BoW concept counts are filtered to events strictly before t^*_f. For cases, t^* is the date of the first qualifying-diagnosis admission (ICD code present). For controls, t^* is an anchor admission chosen to match the case cohort's clinical-history depth at the time of selection. More precisely, for each eligible control admission we count the cumulative OMOP codes recorded up to that admission's feature cutoff, draw a target count from the case cohort's distribution of codes-at-t^*_f, and select the control admission whose code count is closest to that draw. This prevents controls from being trivially distinguishable from cases by data availability (history length) rather than disease-relevant signal. Controls are all MIMIC-IV patients who (a) had at least one admission without the qualifying ICD diagnosis and (b) had at least one prior admission before their matched feature cutoff t^*_f, ensuring MOTOR encodes a non-trivial clinical history. This matching is performed at the cohort level rather than 1:1 against a specific case. All eligible controls form the unlabeled pool $\mathcal{U}$, and disease prevalence in $\mathcal{U}$ reflects natural MIMIC-IV frequencies. Our downstream task is binary classification at a fixed t^*, not a survival-analysis task; there is no follow-up window over which censoring could occur.

B.2. AUPRC normalization

Raw AUPRC is bounded below by prevalence π: a random classifier achieves AUPRC $\approx \pi$, so a raw value of e.g. 0.5 is excellent for a 2.7%-prevalence task but poor for an 83%-prevalence one, making cross-task comparison of raw AUPRC misleading. We define $\text{AUPRC}_\text{n} = \text{AUPRC}/\pi$, so that AUPRC_n=1.0 corresponds to chance-level positive-case ranking and AUPRC_n>1.0 indicates better-than-chance recovery of positive cases, consistent with the prevalence-corrected framing of precision-recall analysis in Davis & Goadrich (2006).

B.3. Model Training Details

Logistic regression uses L2 regularization with $C = 1.0$ (sklearn default). The MLP has two hidden layers of width 128 with ReLU activations, trained with Adam (lr $= 10^{-3}$, max 200 epochs, early stopping on validation loss with patience 10).

XGBoost uses 100 trees, max depth 4, learning rate 0.1, subsample 0.8. All classifiers are retrained from scratch on the updated labeled set at every AL round; no warm-starting is used.

The seed set of 20 patients is drawn with stratified sampling (at least 1 positive, at least 1 negative per fold) to ensure a well-defined binary classifier from round 0. All five folds share the same pool split but different seed draws.

BoW features count per-patient occurrences of condition, drug, measurement, and procedure OMOP concepts appearing strictly before the feature cutoff $t_f^* = t^* - 90$ days, consistent with the gap applied to MOTOR. Concepts appearing in fewer than 5 patients in the training split are discarded; the resulting sparse count matrix is ℓ_2-normalized per patient. Feature dimensionality ranges from $\sim$3,000 to $\sim$15,000 depending on the task.

C. Per-Disease AULC Results

Table 3 reports AULC (AUPRC_n) for each of the 18 disease tasks, as median [IQR] over 5 folds (classifiers averaged), providing per-task uncertainty beyond the cross-task aggregates reported in the main text. The Ceiling column reports AUPRC_n of each classifier trained on the entire labeled pool (no AL, all available patients), normalized by the same per-fold test-set prevalence used throughout the paper and averaged over classifiers and folds, contextualizing how much headroom remains for each task.

Table 3. AULC (AUPRC_n) per disease task (MOTOR embeddings): median [IQR] over 5 folds, classifiers averaged. Tasks sorted by prevalence (rarest first).

Task	Prev.	Random	Entropy	Coreset	K-Means	TypiClust	Ceiling
Lymphoma	0.013	2.85 [2.39–6.92]	6.03 [4.86–8.87]	2.70 [2.61–4.10]	2.75 [1.54–3.49]	2.48 [1.69–4.21]	9.21
Leukemia	0.016	5.39 [5.19–9.10]	14.38 [13.02–14.52]	5.54 [5.40–10.08]	3.28 [3.20–3.31]	5.40 [4.81–6.86]	12.37
Breast Cancer	0.017	2.87 [1.82–2.97]	3.75 [3.13–3.82]	2.77 [1.90–3.13]	1.92 [1.87–2.40]	3.18 [1.93–3.33]	5.25
Prostate Cancer	0.019	2.77 [2.43–2.94]	3.40 [2.61–3.89]	2.63 [2.54–2.80]	2.23 [2.05–3.10]	2.58 [1.89–2.94]	4.20
NAFLD	0.022	2.50 [2.49–3.08]	2.79 [2.72–3.48]	2.35 [2.34–3.08]	2.03 [1.82–2.13]	2.23 [2.22–2.37]	4.26
Lung Cancer	0.024	1.49 [1.43–2.13]	1.89 [1.74–2.54]	1.69 [1.36–1.76]	1.49 [1.41–1.59]	1.46 [1.29–1.95]	2.39
Celiac Disease	0.027	5.86 [5.34–6.32]	10.00 [9.72–10.05]	7.52 [7.45–7.58]	5.36 [5.19–5.69]	5.03 [4.30–5.18]	10.54
Colorectal Cancer	0.059	6.70 [6.48–6.92]	8.90 [8.81–8.91]	6.39 [6.07–6.70]	4.85 [4.70–4.99]	6.43 [5.95–6.74]	8.90
Heart Attack	0.102	1.92 [1.81–2.01]	2.40 [2.23–2.46]	1.90 [1.84–1.96]	1.68 [1.64–1.71]	1.79 [1.67–1.91]	2.57
Sepsis	0.119	1.71 [1.58–1.71]	2.07 [2.02–2.26]	1.73 [1.73–1.84]	1.46 [1.42–1.66]	1.70 [1.55–1.74]	2.40
COPD	0.137	2.50 [2.37–2.50]	3.11 [2.98–3.14]	2.59 [2.59–2.63]	2.39 [2.36–2.41]	2.53 [2.38–2.59]	3.39
Dementia	0.168	1.73 [1.71–1.73]	2.03 [2.02–2.03]	1.68 [1.65–1.68]	1.67 [1.59–1.69]	1.73 [1.72–1.76]	2.23
Stroke	0.210	1.61 [1.56–1.62]	1.87 [1.80–1.97]	1.73 [1.60–1.78]	1.46 [1.45–1.47]	1.61 [1.49–1.63]	2.30
Chr. Kidney Dis.	0.290	1.77 [1.76–1.78]	1.90 [1.89–1.94]	1.74 [1.71–1.74]	1.66 [1.66–1.67]	1.71 [1.70–1.74]	2.07
Atrial Fib.	0.291	1.72 [1.64–1.74]	1.90 [1.83–1.93]	1.79 [1.73–1.86]	1.60 [1.57–1.62]	1.63 [1.61–1.67]	2.08
Type 2 Diabetes	0.295	2.25 [2.20–2.25]	2.37 [2.36–2.37]	2.23 [2.21–2.25]	2.18 [2.15–2.18]	2.22 [2.16–2.25]	2.51
Heart Failure	0.636	1.25 [1.25–1.26]	1.27 [1.26–1.27]	1.26 [1.25–1.26]	1.23 [1.23–1.23]	1.24 [1.24–1.24]	1.37
Hypertension	0.829	1.10 [1.10–1.10]	1.10 [1.09–1.10]	1.10 [1.10–1.10]	1.08 [1.08–1.09]	1.09 [1.09–1.09]	1.14

Entropy's terminal AUPRC_n is closer to, or beyond, the full-data ceiling in 48/54 task$\times$classifier combinations (89%; Figure 6). Entropy outright surpasses the ceiling in 21/54 cells, with the margin strongly concentrated among the rarest diseases (Spearman $r{=}{-}0.70$ between prevalence and margin) and led by leukemia (up to $+4.45$).The exceptions to entropy's advantage are all near-ceiling, comparatively high-prevalence tasks (heart failure, hypertension, lung cancer, prostate cancer) where random is marginally closer. This confirms that entropy's gains are not an artifact of task difficulty and that patient slection can have a significant contirbution on disease risk model performance.

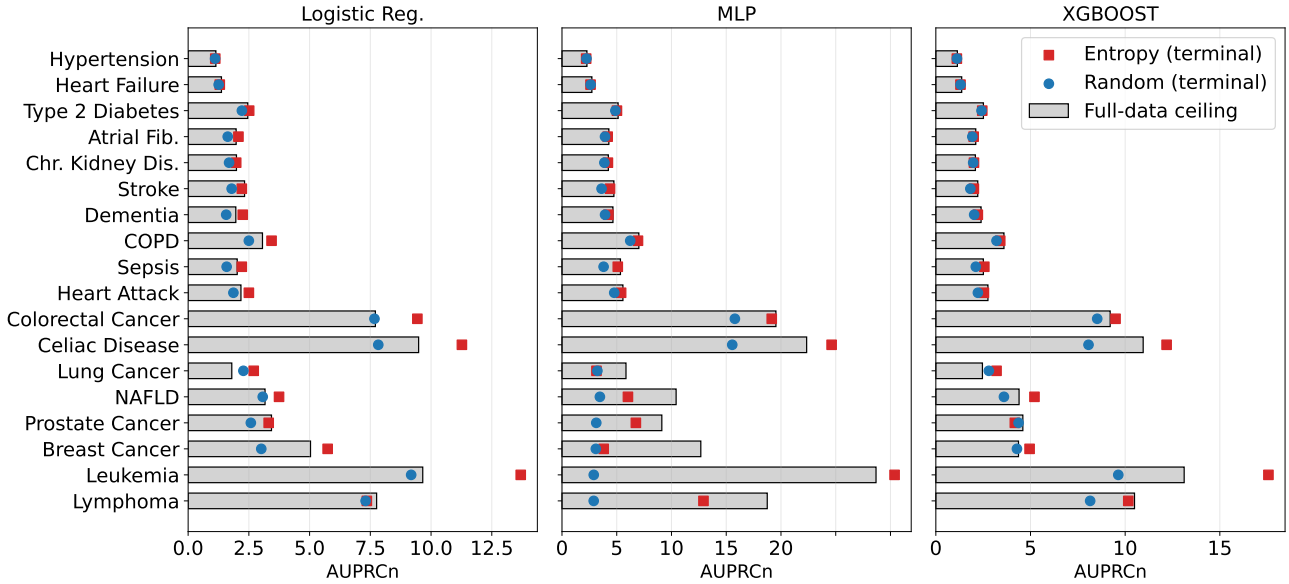


Figure 6. Terminal AUPRC_n (entropy vs. random) against the full-data ceiling (classifier trained on the entire labeled pool, no AL), per task and classifier.

D. Supplementary Figures

D.1. Terminal $AUPRC_n$ per Task

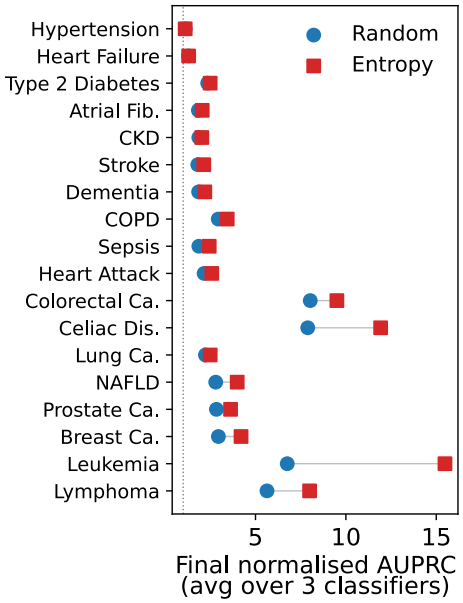


Figure 7. Terminal $AUPRC_n$ per task, sorted by prevalence (rarest top).

D.2. BoW vs. MOTOR Representation

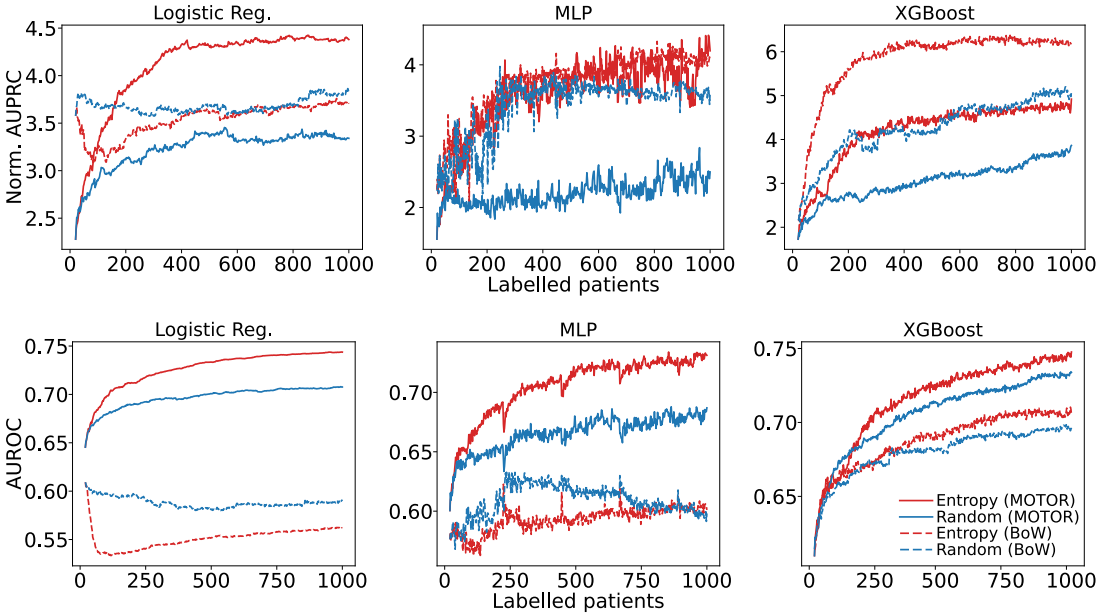


Figure 8. Entropy vs. random on MOTOR (solid) and BoW (dashed). *Top*: $AUPRC_n$. *Bottom*: AUROC.

D.3. AULC Distributions and Learning Curves

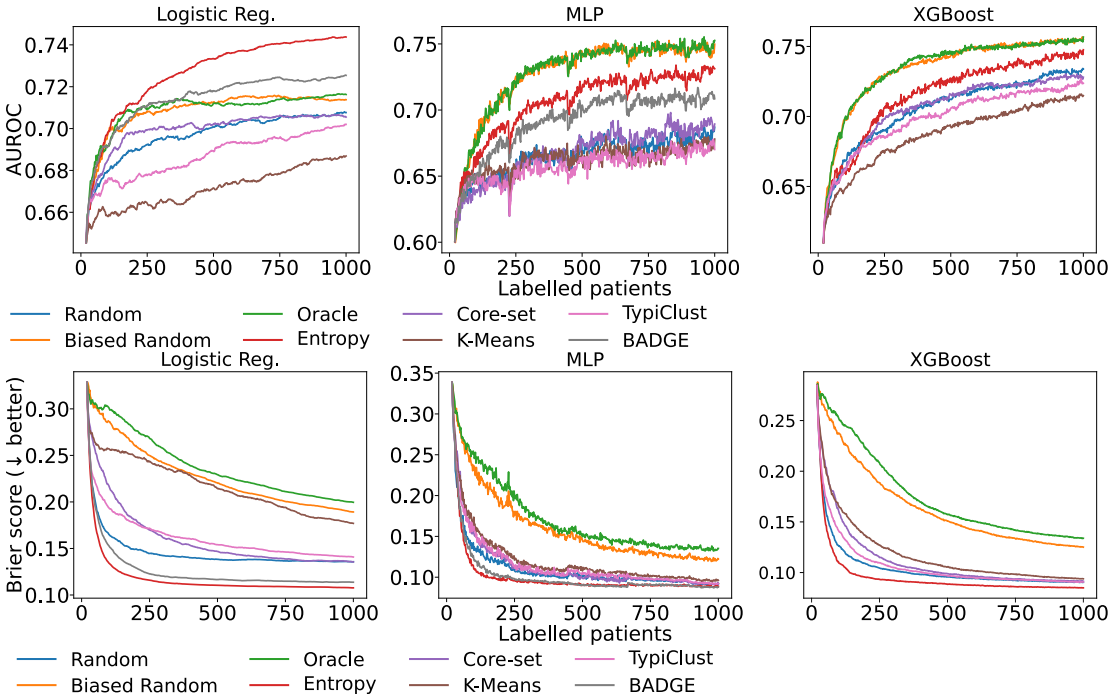
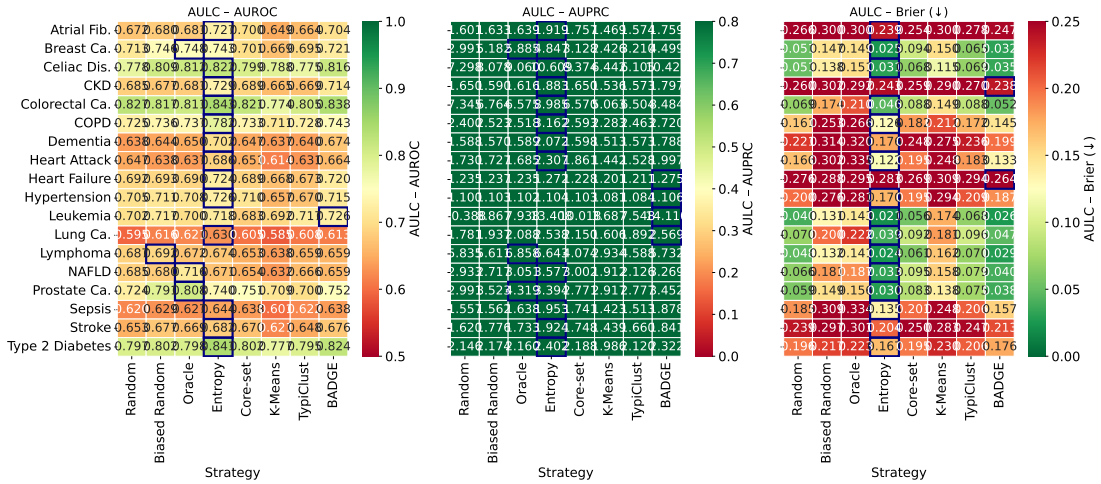


Figure 9. Mean AUROC and Brier learning curves across 18 tasks.

D.4. Per-Task Heatmaps



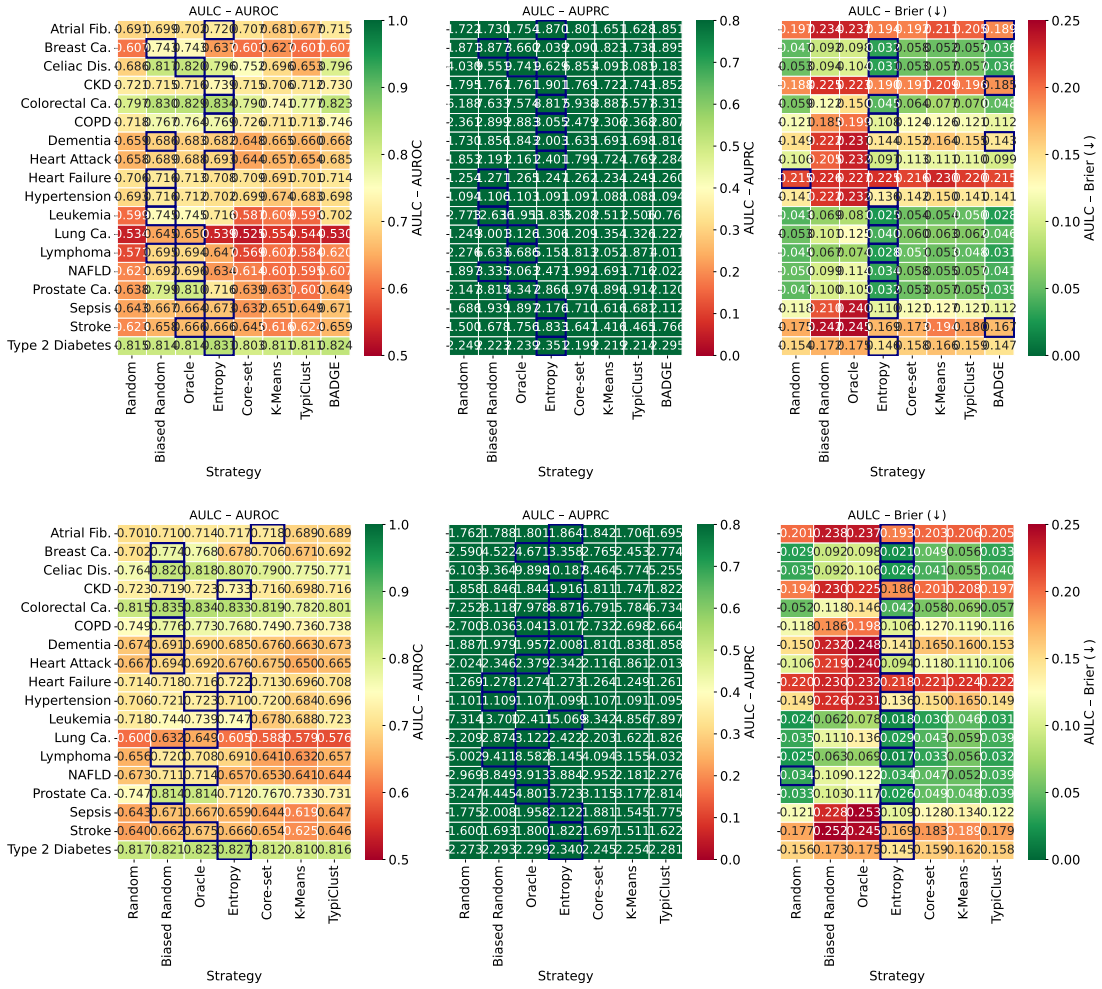
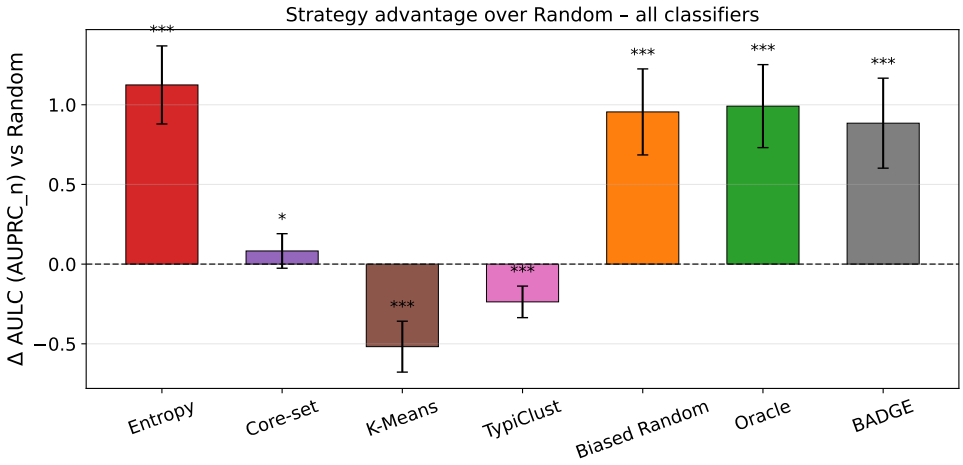


Figure 10. Per-task AULC (AUROC, $AUPRC_n$, Brier) heatmaps. Navy border: best strategy per task. Top: logistic regression. Middle: MLP. Bottom: XGBoost.

D.5. Paired-Difference View



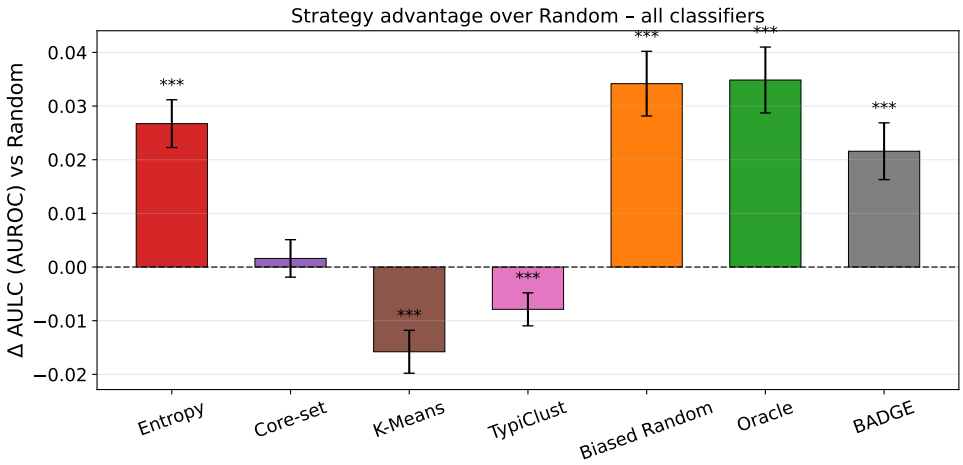


Figure 11. Paired-difference AULC (Δ = Strategy − Random, mean ± 95% CI over task×fold pairs, averaged over classifiers). Top: $AUPRC_n$. Bottom: AUROC. Stars: Wilcoxon signed-rank vs. zero.

E. BADGE Supplementary Experiment

BADGE (Ash et al., 2020) combines uncertainty and diversity via gradient embeddings. As it requires back-propagation, it is incompatible with XGBoost, so we evaluate it on logistic regression and MLP only. Table 4 shows that BADGE performs similarly to entropy, confirming that added geometric diversity does not substantially improve over boundary-proximity selection in this setting.

Table 4. BADGE supplementary results (LR and MLP, MOTOR embeddings, 18 tasks, 5 folds). Mean ± std. BADGE is not evaluated on XGBoost (incompatible gradient requirements).

	Logistic Reg.		MLP	
Strategy	AUROC	$AUPRC_n$	AUROC	$AUPRC_n$
Random	0.697 ±0.066	3.233 ±2.935	0.665 ±0.080	2.204 ±1.258
Entropy	**0.727 ±0.069**	**4.111 ±3.639**	**0.706 ±0.079**	**3.558 ±3.246**
BADGE	0.714 ±0.068	3.946 ±3.702	0.691 ±0.083	3.260 ±2.953